2nd Edition

Data Visualisation

A Handbook for Data Driven Design



Andy Kirk

Data Visualisation

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

2nd Edition

Data Visualisation

A Handbook for Data Driven Design



Andy Kirk

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Aly Owen
Editorial assistant: Lauren Jacobs
Production editor: Ian Antcliff
Copyeditor: Neville Hankins
Proofreader: Christine Bitten
Indexer: David Rudeforth
Marketing manager: Susheel Gokarakonda
Cover design: Shaun Mercier
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in the UK

First edition published 2016. Reprinted four times in 2016, twice in 2017, three times in 2018, and three times in 2019.

Library of Congress Control Number: 2018964578

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-5264-6893-2
ISBN 978-1-5264-6892-5 (pbk)

Contents

Acknowledgements vii
About the Author ix
Discover Your Textbook's Online Resources xi

Introduction 1

PART A FOUNDATIONS 13

1 Defining Data Visualisation 15

2 The Visualisation Design Process 31

PART B THE HIDDEN THINKING 59

3 Formulating Your Brief 61

4 Working With Data 95

5 Establishing Your Editorial Thinking 119

PART C DEVELOPING YOUR DESIGN SOLUTION 133

6 Data Representation 135

7 Interactivity 203

8 Annotation 231

9 Colour 249

10 Composition 277

Epilogue 295
References 301
Index 303

Acknowledgements

I could not have written this book without the unwavering support of my wonderful wife, Ellie, and my family. The book is dedicated to my inspirational Dad who sadly passed away before its publication. I want to acknowledge the contributions of the thousands of data visualisation practitioners who have created such a wealth of exceptional design work and smart writing. I have been devouring this for over a decade now and I am constantly inspired by the talents and minds behind it all. I also want to express my gratitude to the people and organisations who have granted me permission to reference and showcase their visualisation work in this book. Sincere thanks to the many people at Sage who have played a role in making this book grow from the first proposal and now to a second edition. Finally, to you the readers, I am hugely thankful that you chose to invest in this book. I hope it helps you in your journey to learning about this super subject.

About the Author

Andy Kirk is a freelance data visualisation specialist based in Yorkshire, UK. He is a visualisation design consultant, training provider, teacher, author, speaker, researcher and editor of the award-winning website visualisingdata.com.

After graduating from Lancaster University in 1999 with a BSc (hons) in Operational Research, Andy's working life began with a variety of business analysis and information management roles at organisations including CIS Insurance, West Yorkshire Police and the University of Leeds.

He discovered *data visualisation* in early 2007, when it was lurking somewhat on the fringes of the Web. Fortunately, the timing of this discovery coincided with his shaping of his Master's (MA) degree research proposal, a self-directed research programme that gave him the opportunity to unlock and secure his passion for the subject.

He launched visualisingdata.com to continue the process of discovery and to chart the course of the increasing popularity of the subject. Over time, this award-winning site has grown to become a popular reference for followers of the field, offering contemporary discourse, design techniques and vast collections of visualisation examples and resources.

Andy became a freelance professional in 2011. Since then he has been fortunate to work with a diverse range of clients across the world, including organisations such as Google, CERN, Electronic Arts, the EU Council, Hershey and McKinsey. At the time of publication, he will have delivered over 270 public and private training events in 25 different countries, reaching more than 6000 delegates. Alongside his busy training schedule, Andy also provides design consultancy, his primary client being the Arsenal FC Performance Team, since 2015.

In addition to his commercial activities, he maintains regular engagements in academia. Between 2014 and 2015 he was an external consultant on a research project called 'Seeing Data', funded by the Arts & Humanities Research Council and hosted by the University of Sheffield. This study explored the issues of data visualisation literacy among the general public and, inter alia, helped to shape an understanding of the human factors that affect visualisation literacy and the effectiveness of design.

Andy joined the highly respected Maryland Institute College of Art (MICA) as a visiting lecturer in 2013 teaching a module on the Information Visualisation Master's Programme through to 2017. From January 2016, he taught a data visualisation module as part of the MSc in Business Analytics at the Imperial College Business School in London through to 2018. As of May 2019, Andy has started teaching at University College London (UCL).

Discover Your Textbook's Online Resources

Want more support around understanding and creating data visualisations? Andy Kirk is here to help, offline and on!

Hosted by the author and with resources organized by chapter, the supporting website for this book has everything you need to explore, practice, and hone your data visualisation skills.

- **Explore the field**: expand your knowledge and reinforce your learning about working with data through libraries of further reading, references, and tutorials.
- **Try this yourself**: revise, reflect, and refine your skill and understanding about the challenges of working with data through practical exercises.
- **See data visualisation in action**: get to grips with the nuances and intricacies of working with data in the real world by navigating instalments of the narrative case study and seeing an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.
- **Chartmaker directory**: access crowd-sourced guidance that aims to answer the crucial question 'which tools make which charts?' with this growing directory of examples and technical solutions for chart building.

Ready to learn more? Go beyond the book and dive deeper into data visualisation via the rest of Andy's website (**www.visualisingdata.com**), which contains **data visualisation tools and software**, links to additional influential further reading, and a **blog with monthly collections** of the best data visualisation examples and resources each month.

Introduction

The primary challenge one faces when writing a book about data visualisation is to determine what to leave in and what to leave out. Data visualisation is a big subject. There is no single book to rule it all because there is no one book that can truly cover it all. Each and every one of the topics covered by the chapters in this book could (and, in several cases, do) exist as books in their own right.

The secondary challenge when writing a book about data visualisation is to decide how to weave the content together. Data visualisation is not rocket science; it is not an especially complicated discipline, though it can be when working on sophisticated topics and with advanced applications. It is, however, a complex subject. There are lots of things to think about, many things to do and, of course, things that will need making. Creative and journalistic sensibilities need to blend harmoniously with analytical and scientific judgement. In one moment, you might be checking the statistical rigour of an intricate calculation, in the next deciding which shade of orange most strikingly contrasts with a vibrant blue. The complexity of data visualisation manifests in how the myriad small ingredients interact, influence and intersect to form a whole.

The decisions I have made when formulating this book's content have been shaped by my own process of learning. I have been researching, writing about and practising data visualisation for over a decade. I believe you only truly learn about your own knowledge of a subject when you have to explain it and teach it to others. To this extent I have been fortunate to have had extensive experience designing and delivering commercial training as well as academic teaching.

I believe this book offers an effective and proven pedagogy that successfully translates the complexities of this subject in a form that is fundamentally useful. I feel well placed to bridge the gap between the *everyday* practitioners, who might identify themselves as beginners, and the superstar talents expanding the potential of data visualisation. I am not going to claim to belong to the latter cohort, but I have certainly been a novice, taking tentative early steps into this world. Most of my working hours are spent helping others start their journey. I know what I would have valued when I started out in this field and this helps inform how I now pass this on to others in the same position I was several years ago.

There is a large and growing library of fantastic books offering different theoretical and practical viewpoints on this subject. My aim is to add value to this existing collection by approaching the subject through the perspective of process. I believe the path to mastering data visualisation is achieved by making better decisions: namely, effective choices, efficiently made. I will help you understand what decisions need to be made and give you the confidence to make the right choices. Before moving on to discuss the book's intended audience, here are its key aims:

- To **challenge** your existing approaches to creating and consuming visualisations. I will challenge your beliefs about what you consider to be effective or ineffective visualisation. I will encourage you to eliminate arbitrary choices from your thinking, rely less on taste and instinct, and become more reasoned in your judgements.
- To **enlighten** you I will increase your awareness of the possible approaches to visualising data. This book will broaden your visual vocabulary, giving you a wider and more sophisticated understanding of the contemporary techniques used to express your data visually.
- To **equip** is to provide you with robust tactics for managing your way through the myriad options that exist in data visualisation. To help you overcome the burden of choice, an adaptable framework is offered to help you think for yourself, rather than relying on inflexible rules and narrow instruction.
- To **inspire** is to open the door to a subject that will stimulate you to elevate your ambition and broaden your confidence. Developing competency in data visualisation will take time and will need more than just reading this book. It will require a commitment to embrace the obstacles that each new data visualisation opportunity poses through practice. It will require persistence to learn, apply, reflect and improve.

Who Is This Book Aimed At?

Anyone who has reason to use quantitative and qualitative methods in their professional or academic duties will need to grasp the demands of data visualisation. Whether this is a large part of your duties or just a small part, this book will support your needs.

The primary intended audiences are undergraduates, postgraduates and early-career researchers. Although aimed at those in the social sciences, the content will be relevant to readers from across the spectrum of arts and humanities right through to the natural sciences.

This book is intended to offer an accessible route for novices to start their data visualisation learning journey and, for those already familiar with the basics, the content will hopefully contribute to refining their capabilities. It is not aimed at experienced or established visualisation practitioners, though there may be some new perspectives to enrich their thinking: some content will reinforce existing knowledge, other content might challenge their convictions.

The people who are active in this field come from all backgrounds. Outside academia, data visualisation has reached the mainstream consciousness in professional and commercial contexts. An increasing number of professionals and organisations, across all industry types and sizes, are embracing the importance of getting more value from their data and doing more with it, for both internal and external benefit. You might be a market researcher, a librarian or a data analyst looking to enhance your data capabilities. Perhaps you are a skilled graphic designer or web developer looking to take your portfolio of work into a more data-driven direction. Maybe you are in a managerial position and though not directly involved in the creation of visualisation work, you might wish to improve the sophistication of the language you coordinate or commission others who are. Everyone needs the lens and vocabulary to evaluate work effectively.

Data visualisation is a genuinely multidisciplinary discipline. Nobody arrives fully formed with all constituent capabilities. The pre-existing knowledge, skills or experiences which, I think, reflect the traits needed to get the most out of this book would include:

- Strong numeracy is necessary as well as a familiarity with basic **statistics**.
- While it is reasonable to assume limited prior knowledge of data visualisation, there should be a strong **desire** to want to learn it. The demands of learning a craft like this take time and effort; the capabilities will need nurturing through ongoing learning and practice. They are not going to be achieved overnight or acquired alone from reading this book. Any book that claims to be able magically to inject mastery through just reading it cover to cover is over-promising and likely to under-deliver.
- The best data visualisers possess inherent **curiosity**. You should be the type of person who is naturally disposed to question the world around them. Your instinct for discovering and sharing answers will be at the heart of this activity.
- There are no expectations of your having any prior familiarity with **design** principles, but an appetite to embrace some of the creative aspects presented in this book will heighten the impact of your work. Time to unleash that suppressed imagination!
- If you *are* somebody fortunate to possess already a strong creative flair, this book will guide you through when and crucially when *not* to tap into this sensibility. You should be willing to increase the rigour of your **analytical** decision making and be prepared to have your creative thinking informed more fundamentally by data rather than just instinct.
- No particular **technical** skills are required to get value from this book, as I will explain shortly. But you will ideally have some basic knowledge of spreadsheets and experience of working with data irrespective of which particular tool.

This is a portable practice involving techniques that are subject-matter agnostic. Throughout this book you will see a broad array of examples from different industries covering many different topics. Do not be deterred by any example being about a subject different to your own area of interest. Look beyond the subject and you will see analytical and design choices that are just as applicable to you and your work: a line chart showing political forecasts involves the same thought process as would a line chart showing stock prices changing or average global temperatures rising. A line chart is a line chart, regardless of the subject matter.

The type of data you are working with is the only legitimate restriction to the design methods you might employ, not your subject and certainly not traditions in your subject. 'Waterfall charts are only for people in finance', 'maps are only for cartographers', 'Sankey diagrams are only for engineers'. Enter this subject with an open mind, forget what you believe or have been told is the *normal* approach, and your capabilities will be expanded.

Data visualisation is an entirely global community, not the preserve of any geographic region. Although the English language dominates written discourse, the interest in the subject and work created from studios through to graphics teams originates everywhere. There are cultural influences and different flavours in design sensibility around the world which enrich the field but, otherwise, it is a practice common and accessible to all.

Finding the Balance

Handbook vs Manual

The description of this book as a 'handbook' positions it as distinct from a tutorial-based manual. It aims to offer conceptual and practical guidance, rather than technical instruction. Think of it more as a guidebook for a tourist visiting a city than an instruction manual for how to fix a washing machine.

Apart from a small proportion of visualisation work that is created manually, the reliance on technology to create visualisation work is an inseparable necessity. For many beginners in visualisation there is an understandable appetite for step-by-step tutorials that help them immediately to implement their newly acquired techniques.

However, writing about data visualisation through the lens of selected tools is hard, given the diversity of technical options that exist in the context of such varied skills, access and needs. The visualisation technology space is characterised by flux. New tools are constantly emerging to supplement the many that already exist. Some are proprietary, others are open source; some are easier to learn but do not offer much functionality; others do offer rich potential but require a great deal of foundation understanding before you even accomplish your first bar chart. Some tools evolve to keep up with current techniques; they are well supported by vendors and have thriving user communities, others less so. Some will exist as long-term options whereas others depreciate. Many have briefly burnt brightly but quickly become obsolete or have been swallowed up by others higher up the food chain. Tools come and go but the craft remains.

There is a role for all book types and a need for more than one to acquire true competency in a subject. Different people want different sources of insight at different stages in their development. If you *are* seeking a text that provides instructive tutorials, you will learn from this how to accomplish technical developments in a given technology. However, if you *only* read tutorial-based books, you will likely fall short in the fundamental critical thinking that will be needed to harness data visualisation as a skill.

I believe a practical, rather than technical, text focusing on the underlying craft of data visualisation through a tool-agnostic approach offers the most effective guide to help people learn this subject.

The content of this book will be relevant to readers regardless of their technical knowledge and experience. The focus will be to take your critical thinking towards a detailed, fully reasoned design specification – a declaration of intent of what you want to develop. Think of the distinction as similar to that between architecture (design specification) and engineering (design execution).

There is a section in Chapter 3 that describes the influence technology has on your work and the places it will shape your ambitions. Furthermore, among the digital resources offered online are further profiles of applications, tools and libraries in common use in the field today and a vast directory of resources offering instructive tutorials. These will help you to apply technically the critical capabilities you acquire throughout this book.

Useful vs Beautiful

Another important distinction to make is that this book is not intended to be seen as a beauty pageant. I love flicking through glossy 'coffee table' books as they offer great inspiration, but often lack substance beyond the evident beauty. This book serves a different purpose to that. I believe, for a beginner or relative beginner, the most valuable inspiration comes more from understanding the thinking behind some of the amazing works encountered today, learning about the decisions that led to their conceptual development.

My desire is to make this the most *useful* text available, a reference that will spend more time on your desk than on your bookshelf. To be useful is to be used. I want the pages to be dog-eared. I want to see scribbles and annotated notes made across its pages and key passages underlined. I want to see sticky labels peering out above identified pages of note. I want to see creases where pages have been folded back or a double-page spread that has been weighed down to keep it open. It will be an elegantly presented and packaged book, but it should not be something that invites you to look but not touch.

Pragmatic vs Theoretical

The content of this book has been formed through years of absorbing knowledge from as many books as my shelves can hold, generations of academic work, endless web articles, hundreds of conference talks, personal interactions with the great and the good of the field, and lots and lots of practice. More accurately, lots and lots of mistakes. What I present here is a pragmatic distillation of what I have learned and feel others will benefit from learning too.

It is not a deeply academic or theoretical book. Experienced or especially curious practitioners may have a desire for deeper theoretical discourse, but that is beyond the intent of this particular text. You have to draw a line somewhere to determine the depth you can reasonably explore about a given topic. Take the science of visual perception, for example, arguably the subject's foundation. There is no value in replicating or attempting to better what has already been covered by other books in greater quality than I could achieve.

An important reason for giving greater weight to pragmatism is because of the inherent imperfections of this subject. Although there is so much important empirical thinking in this subject, the practical application can sometimes fail to translate beyond the somewhat artificial context of a research study. Real-world circumstances and the strong influence of human factors can easily distort the significance of otherwise robust concepts.

Critical thinking will be the watchword, equipping you with the independence of thought to decide rationally for yourself which solutions best fit your context, your data, your message and your audience. To accomplish this, you will need to develop an appreciation of all the options available to you (the different things you *could* do) and a reliable approach for critically determining what choices you should make (the things you *will* do and *why*).

Contemporary vs Historical

I have huge respect for the ancestors of this field, the dominant names who, despite primitive means, pioneered new concepts in the visual display of statistics to shape the foundations of the field being practised today. The field's lineage is decorated by pioneers such as William Playfair, W. E. B. Du Bois, Florence Nightingale and John Snow, to name but a few. To many beginners in the field, the historical context of this subject is of huge interest. However, this kind of content has already been covered by plenty of other book and article authors.

I do not want to bloat this book with the unnecessary reprising of topics that have been covered at length elsewhere. I am not going to spend time attempting to enlighten you about how we live in the age of 'Big Data' and how occupations related to data are or will be the 'sexiest jobs' of our time. The former is no longer news, the latter claim emerged from a single source. There is more valuable and useful content I want you to focus your time on.

The subject matter, the ideas and the practices presented here will hopefully not date a great deal. Of course, many of the graphic examples included in the book will be surpassed by newer work demonstrating similar concepts as the field continues to develop. However, their worth as exhibits of a particular perspective covered in the text should prove timeless. As time passes there will be new techniques, new concepts and new, empirically evidenced rules. There will be new thought-leaders, new sources of reference and new visualisers to draw insight from. Things that prove a manual burden now may become seamlessly automated in the near future. That is the nature of a fast-growing field.

Analysis vs Communication

A further distinction to make concerns the subtle but critical difference between visualisation used for analysing data and visualisation used for communicating data.

Before a visualiser can confidently decide what to communicate to others, he or she needs to have developed an intimate understanding of the qualities and potential of the data. In certain contexts, this might only be achieved through exploratory data analysis. Here, the visualiser and the viewer are the same person. Through visual exploration, interrogations of the data can be conducted to learn about its qualities and to unearth confirmatory or enlightening discoveries about what insights exist.

Visualisation for analysis is part of the journey towards creating visualisation for communication, but the techniques used for visual analysis do not have to be visually polished or necessarily appealing. They are only serving the purpose of helping you truly to learn about your data. When a data visualisation is being created to communicate to others, many careful considerations come into play about the requirements and interests of the intended audience. This influences many design decisions that do not exist alone with visual analysis.

For the scope of this book the content is weighted more towards methods and concerns about communicating data visually to others. If your role is concerned more with techniques for

exploratory analysis rather than visual communication, you will likely require a deeper treatment of the topic than this book can reasonably offer.

Another matter to touch on here concerns the coverage of statistics, or lack thereof. For many people, statistics can be a difficult topic to grasp. Even for those who are relatively numerate and comfortable working with simple statistical methods, it is quite easy to become rusty without frequent practice. The fear of making errors with intricate statistical calculations depresses confidence and a vicious circle begins.

You cannot avoid the need to use *some* statistical techniques if you are going to work with data. I will describe some of the most relevant statistical techniques in Chapter 4, at the point in your thinking where they are most applicable. However, I do believe the range and level of statistical techniques *most* people will need to employ on *most* of their visualisation tasks can be overstated. I know there will be exceptions, and a significant minority will be exposed to requiring advanced statistical thinking in their work.

It all depends, of course. In my experience, however, the majority of data visualisation challenges will generally involve relatively straightforward *univariate* and *bivariate* statistical techniques to describe data. Univariate techniques help you to understand the shape, size and range of a single variable of data, such as determining the minimum, maximum and average height of a group of people. Bivariate techniques are used to observe possible relationships between two different variables. For example, you might look at the relationship between gross domestic product and medal success for countries competing at the Olympics. You may also encounter visualisation challenges that require a basic understanding of probabilities to assist with forecasting risk or modelling uncertainty.

The more advanced applications of statistics will be required when working with larger complicated datasets, where *multivariate* techniques are employed simultaneously to model the significance of relationships between multiple variables. Above and beyond that, you are moving towards advanced statistical modelling and algorithm design.

Though it may seem unsatisfactory to offer little coverage of this topic, there is no value in reinventing the wheel. There are hundreds of existing books better placed to offer the depth you might need. That statistics is such a prolific and vast field in itself further demonstrates how deeply multidisciplinary a field visualisation truly is.

Chapter Contents

The book is organised into three main parts (A, B and C) comprising ten chapters and an Epilogue. Each chapter opens with a preview of the content to be covered and closes with a summary of the most salient learning points to emerge. There are collections of further resources available online to substantiate the learning from each chapter.

For most readers, especially beginners, it is recommended that you start from the beginning and proceed through each chapter as presented. For those setting out to begin working on their own visualisation, you might jump straight into Chapters 2–5 to ensure you are fully prepared

for some of the important preparatory activities you need to accomplish before moving on to look at developing your design solution. For those with more experience and/or prior exposure to this subject, who are perhaps looking to fine-tune specific aspects of their design skills, most of your interest will lie in Part C, comprising Chapters 6–10. For readers who just want to dip in and out of specific topic areas, although each chapter builds sequentially from the preceding ones, they can all be read in isolation. Follow any sequence that satisfies your needs. The coloured tabs on the outer edge will provide quick visual navigation through the distinct parts and chapters within.

Part A: Foundations

Part A introduces some important foundational understanding about data visualisation as a subject area and as an activity. The contents of the first two chapters give shape to the coverage across the rest of the book.

Chapter 1 'Defining Data Visualisation' will be the logical starting point for those who are new to the field, providing a definition for the subject and exploring some of the tensions that enrich this subject. The second section explains some of the distinctions and overlaps with other related disciplines. If you already know what data visualisation is about, you might choose to pass on this; it does, though, help frame many of the discussions elsewhere.

Chapter 2 'The Visualisation Design Process' introduces the value of following a design process, the sequence of activities around which the book's contents in Parts B and C are organised. It explains what is involved and offers some useful tips to help you seamlessly adopt this approach. Where the process offers organisation and efficiency, design principles ensure effectiveness. The second section will describe what separates the good from the bad in visualisation design, building up your convictions to help with your upcoming decision making.

Part B: The Hidden Thinking

Part B profiles the first three stages of the data visualisation design process. These are the hidden preparatory stages that will significantly influence the path you take towards an eventual solution.

Chapter 3 'Formulating Your Brief' covers the opening tasks involved in initiating, defining and planning the requirements of your work. The first section looks at issues around context, specifically about the importance of defining curiosity and identifying the circumstances that will shape your project. The second section considers the vision of your work, looking at what purpose it intends to serve and how you might creatively define the type of work you will need to pursue. Finally, a short section looks at the value of harnessing initial ideas.

Chapter 4 'Working With Data' commences your practical involvement with your data, stepping through the four distinct steps that acquaint you with the potential of your

critical raw material. Data acquisition outlines the different origins of and methods for obtaining your data. Data examination profiles the different characteristics that define the type, extent and condition of your data. Data transformation builds on your examination work to find ways of modifying and enhancing your data to prepare it for use. Finally, data exploration discusses methods for discovering more about the qualities and insights hidden away in your data.

Chapter 5 'Establishing Your Editorial Thinking' reflects on the possibilities offered by your data and explains the importance of committing to an editorial path. The chapter opens with a definition about the influence of editorial thinking, using two case studies to explain how editorial definitions influence design choices later in the process.

Part C: Developing Your Design Solution

Part C represents the main part of this book and covers the five distinct layers of the data visualisation anatomy. They are presented in separate chapters to help organise your thinking and to avoid being overwhelmed by the detailed options that exist. However, they are ultimately interrelated matters and the chapter sequencing across this part is carefully arranged to support this. Each chapter follows a similar structure, opening with an array of different possible design options and supplemented by guidance on the factors that will influence your choices. Initially, you will need to make decisions about what elements to include around data representation (charts), interactivity and annotation. You will then complete your thinking about the appearance of these elements, through colour and composition.

Chapter 6 'Data Representation' introduces the act of visual encoding and then expands on this to provide a detailed profile of 49 distinct chart types to help broaden your visual vocabulary. The chapter closes with a run through the key factors that will influence the suitability of your data representation choices.

Chapter 7 'Interactivity' introduces the potential value of incorporating interactive features in your work, profiling a wide range of options – such as filtering, highlighting and animating – that will enable users to interrogate and control a visualisation. The chapter closes with the main considerations that will influence your selection of interactive features.

Chapter 8 'Annotation' describes the importance of providing useful assistance to your viewers, including headings, chart apparatus, and labels. The chapter closes with a look at which factors will inform the choices you make.

Chapter 9 'Colour' commences with an overview of different colour models. This provides the basis for understanding the different ways of applying colour to facilitate data legibility and deliver functional decoration. Once again, having introduced the options, we will look at how you arrive at appropriate choices.

Chapter 10 'Composition' explores the final element of developing your design solution concerning how you organise the placement and sizing of all your visual elements within the space you have to work. Looking at matters of layout, arrangement and chart sizing, we will then wrap up this topic with a discussion about how to make your decisions.

Epilogue: To close the book, the epilogue will summarise the development cycle of activities you will need to undertake as you move your detailed design specification to a fully executed solution.

Digital Resources

The opportunity to supplement the print version of this book with further digital companion resources helps to offer readers a range of additional learning materials:

- a written and video-based case-study of a visualisation project that demonstrates the design process in action;
- an extensive and up-to-date catalogue of over 350 data visualisation tools;
- a large collection of tutorials and resources to help develop your technical capabilities in making a wide range of different charts;
- useful exercises designed to help embed the learning covered in each chapter;
- a digital gallery of all the artwork included in this book and many further examples of the concepts presented across all chapters;
- refreshed reading resources to support ongoing learning about the subjects covered in each chapter.

Glossary

Consistency in the meaning of language and terms used in data visualisation is important. Though data visualisation is no different to many fields that get bogged down by superfluous semantic noise, it can only help to establish clarity about its usage in this book at least.

Roles

Visualiser: This is the role I am assigning to you – the person making the visualisation. Sometimes people prefer to use terms like researcher, analyst, developer, storyteller or even ‘visualist’. Designer would also be particularly appropriate, but I want to broaden the scope of the role beyond just design to cover all activities involved in this discipline.

Viewer: This is the role assigned to the recipient, who is viewing or using your visualisation product. It offers a broader and better fit than alternatives such as consumer, reader, user or customer. However, ‘user’ will be temporarily adopted during the more active chapter about interactivity.

Audience: This concerns the collective group of viewers for whom your work is intended. Within an audience there will be cohorts of different viewer types that you might characterise through distinct personas to help your thinking about serving their varied needs.

Consuming: This will be the general act of the viewer, to consume. I will use more active descriptions like 'reading' and 'using' when consuming becomes too passive or vague, and when distinctions are needed between reading a chart and using interactive features.

Data

Raw data: For the purpose of this book, raw data will be the initial state of data you have collected, received or downloaded that has not yet been subjected to any statistical or transforming treatment. Some people take issue with the implied 'rawness' this label implies, given that data will have already lost its raw state having been recorded by some instrument, stored, retrieved and maybe cleaned already. I appreciate this viewpoint but think it is the most pragmatic label relevant to most people's understanding.

Data source: This is the term used to describe the origin(s) of the raw data used in a visualisation.

Dataset: A table of data is an array of values visually arranged into rows and columns, usually existing in a spreadsheet or database. The rows are the records – instances or items – and the columns are the variables – details about the items. Datasets are visualised in order to 'see' the size, patterns and relationships that are otherwise hard to observe. A dataset may comprise one or a collection of several tables.

Tabulation: For the purpose of this book, I distinguish between types of datasets that are 'normalised' and others that are 'cross-tabulated'. This distinction will be explained in context in Chapter 4.

Data types: The variables (columns) in a table that hold details about items (records) will have different scales of measurement or *data types*. At the most general level, distinctions in quantitative (e.g. salary) and categorical (e.g. gender) data are important in how you will statistically and visually handle them. A detailed distinction between data types, with examples, will again be offered in Chapter 4.

Series: A series of values is essentially a sequence of related values in a table. An example of a series would be the highest recorded temperatures in a city for each day over a month. Though individual daily values will be stored as distinct moment-in-time measurements, the activity of temperature never stops 'happening' and therefore the collected values have a legitimate continuous relationship through the series.

Visualisation

Project: For the purpose of this book, we will consider the development of a data visualisation as being a *project*. Even though you might consider something a quick, small task, it will still need to involve the thinking consistent with the stages of the process covered in this book.

Chart type: Charts are visual representations of data. There are many ways of representing your data, using different combinations of marks, attributes, layouts and apparatus.

Their combinations form archetypes of charts more commonly named *chart types*, such as the bar chart, dendrogram or treemap.

Graphs, plots, diagrams and maps: Traditionally the term *graph* has been used to describe visualisations that display network relationships, while *chart* would be commonly used to label common devices like the bar or pie chart. *Plots* and *diagrams* are more specifically attached to special types of displays but with no pattern of consistency in their usage. All these terms are so interchangeable that any energy expended in explaining meaningful difference is redundant. For the purpose of this book, I will generally stick to the term *chart* to act as the main label to cover all representation types. In places, this 'umbrella' term will incorporate thematic maps, for the sake of convenience, even though they clearly have a visual structure that is quite different to standard charts.

Graphic: The term *graphic* will be used when referring to visuals more focused on information-led displays such as explanation or process diagrams as distinct from charts that are concerned with data-driven visuals. It might also be used to refer more broadly to a visualisation that incorporates charts, text and images.

Format: This concerns the difference in output form between printed work, digital work and physical visualisation work.

Functionality: This concerns the difference in whether a visualisation is static or interactive. Interactive visualisations allow you to manipulate and interrogate a computer-based display of data. They are published on the Web, exist within apps, or are on larger digital displays, as in galleries. In contrast, a static visualisation displays a non-changeable, still display of data that could be published in print but also digitally. Just because something is published digitally does not automatically make it interactive.

Axes: Many common chart types have axis lines that provide a reference for measuring quantitative values or positioning categorical values. The horizontal axis is known as the x-axis and the vertical axis is known as the y-axis.

Scales: Scales exist in two forms, typically. Firstly, as a set of marks along an axis that indicate positions for the range of values included in a chart. Scales are normally presented in regular intervals (10, 20, 30, etc.) representing units of measurement, such as prices, distances, years or percentages. A scale may also be presented in a key to explain associations between, for example, different sizes of areas or classifications of different colour attributes.

Legend: Charts that employ visual attributes, such as colours, shapes or sizes to represent values of data, will often be accompanied by a legend to house visual explanations of classifications, known as keys.

Outliers: Outliers are points of data that are outside the normal range of values. They are the unusually large or small or simply different values that stand out and generally draw a viewer's attention.

Correlation: This is a measure of the presence and extent of a mutual relationship between two or more variables of data. For example, you would expect to see a correlation between the height and weight of people or age and salary of workers. Devices like scatter plots, in particular, help visually to portray possible correlations between two quantitative values.

Part A

Foundations



1

Defining Data Visualisation

This opening chapter will introduce data visualisation through the prism of a proposed definition. Each component that forms this definition will be explored in depth to illustrate some of the main characteristics and complexities of this subject.

The second part of the chapter will position data visualisation in the context of other related disciplines or fields, explaining where overlaps or clear distinctions exist. Overall, this chapter will seek to forge a shared understanding that will help set the tone and reasoning for the structure of this book.

1.1 What Is Data Visualisation?

It is useful to commence this book with a definition of data visualisation (Figure 1.1). It helps to ensure we (you the *reader*, me the *writer*) have a mutual understanding, from the outset, about what is meant by data visualisation in the context of this text. The components of this definition carve the subject into distinct perspectives around which the contents of this book are organised.

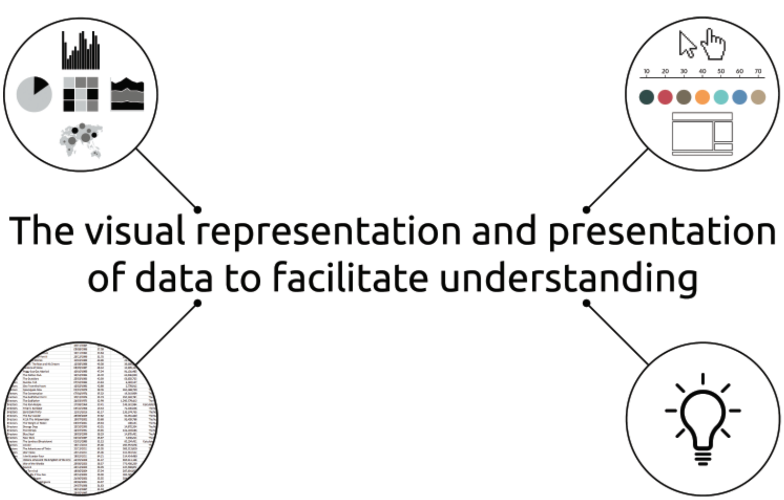


Figure 1.1 A Definition for Data Visualisation

Let me delve into this and describe the roles of and relationships between each component expressed. I will also explain where and how these topics will be covered. Firstly, let's look at **data**.

Data is names and amounts. It is groupings, descriptions and measurements. It is dates and locations. It will be helpful for discussions in this book to think of data as being typically structured in table form, with rows of records and columns of variables. Most data we commonly encounter will exist in textual, numeric or a combined form, but it is also worth noting the opportunities that increasingly exist for working with data assets in media forms of images, audio and video.

In Chapter 4 you will learn about the importance of developing an intimate understanding of your data to acquaint yourself fully with its properties, its condition and its qualities.

You will see that data is the fundamental element driving the decisions across this design process. Without data there is no material to feed nor necessitate a visualisation. Conversely, without visualisation the value of data can be unfulfilled. This is not to say we should *always* visualise data, absolutely not, but in most circumstances, to harness the maximum value of data, there are missed opportunities if we do not.

To explain, here is a simple illustration. When data is presented in a table, it is a straightforward task for a viewer to scan the rows and columns to seek out values of relevance or to discover particular data points that trigger interest. For instance, by viewing the table in Figure 1.2 it should prove quite simple to find out what the percentage share of online sales for a Company X was during April 2016. Now look for the percentage share of store sales during December 2011.

Figure 1.2 Proportion of Sales % by Channel Over Time

Proportion of Sales % by Channel Over Time

REPORTING MONTH	STORES	ONLINE	TELEPHONE
May 2011	71	29	0
Jun 2011	72	28	0
Jul 2011	71	28	1
Dec 2011	71	28	1
Jun 2012	73	26	1
Jul 2012	77	22	1
Sep 2012	75	24	1
Nov 2012	75	24	1
Jun 2013	73	26	1
Nov 2013	73	26	1
Jan 2014	73	26	1
Jun 2014	72	27	1
Aug 2014	55	44	1
Sep 2014	60	38	2
Oct 2014	51	48	1
Nov 2014	44	55	1
Jan 2015	52	47	1
Mar 2015	50	48	2
Jun 2015	49	49	2
Jul 2015	37	61	2
Aug 2015	40	58	2
Nov 2015	40	59	1
Dec 2015	22	77	1
Jan 2016	21	77	2
Feb 2016	20	78	2
Apr 2016	14	84	2
Dec 2016	21	77	2
Jun 2018	6	93	1
Jul 2018	6	93	1
Dec 2018	0	100	0

As a viewer your task is simply to find the relevant row and column intersection: look at the value display and read it. The percentage share of online sales for Company X during April 2016 is 84, and for store sales during December 2011 it is 71.

To find which sales channel had the second largest percentage share of sales during August 2014, again just find the relevant row, compare the three quantitative values along that row, and then determine which channel column contains the second-ranked amount. For this month, the online channel, at 44, had the second largest percentage share of sales.

The limitations of reading data when it is presented in this form emerge when we want to answer broader questions: that is, enquiries that transcend the scope of an answer originating from a single or small number of adjacent data points. From the same table, how easy do you find it to identify the headline trends across each sales channel over the period of time displayed?

You can probably ascertain that the percentage share of sales for stores starts quite high then drops to nothing, the percentage share of online sales starts quite low and then reaches the 100% maximum, and the percentage share of sales via telephone is consistently tiny.

Though it takes a while to study the values under each sales channel column in order to form this summary observation, it is still possible. But what if your observations need to be formed more quickly? What if you needed to know more about the localised patterns of ups and downs within those global trends? What if you wanted to identify the first occasion when the percentage share of online sales exceeded the percentage share of store sales? When was the last occasion the percentage share of store sales exceeded that of online sales? During which periods did the different sales channels experience the most accelerated upward or downward changes?

These are harder questions to answer efficiently and accurately from the data alone. This is because synthesising observations from multiple values across different rows and columns to perceive broader relationships fails to exploit fully the capabilities of our visual system – how our eyes and mind work together to make sense of objects and patterns. To read values in isolation, store them in our short-term memory and compare them in our head with other isolated values is mentally challenging. It is not impossible, since we can still accomplish this with just a table of data, but it will take an excessive amount of time and effort.

This workload will also only increase as the data grows in volume and complexity. For instance, what if this table were 1000 rows deep and there were 20, 50 or 100 different columns to work through? Or, what if the quantities had similar value sizes and more modest variation? How easy would it then be to notice significant patterns?

The crux of all this is that we can *look* at data, but we cannot really *see* it. To see data, we need to represent it in a different, visual form.

Returning to the definition, the term **visual representation** is arguably the quintessential activity of data visualisation. Representation involves making decisions about how you are going to portray your data visually so that the subject understanding it offers can be made accessible to your audience. In simple terms, this is all about charts and the act of selecting the right chart to show the features of your data that you think are most relevant.

The building blocks of any chart are *marks* and *attributes*. Marks can be points, lines or shapes and they are used to represent items of data. An example of an item of data from the table in

Figure 1.2 would be the 'percentage share of sales from stores during June 2014'. Not the value itself, more the *thing* the value is about.

Attributes, sometimes described as channels, are visual variations of marks to represent the values associated with each. These include properties such as different scales of size, colour or position. If the item of data is 'percentage share of sales from stores during June 2014', an attribute would be used to represent the associated value, in this case 72. If marks and attributes are the ingredients, the different combinations used create different chart *types* – the recipes.

Figure 1.3 shows a chart of the data shown in the table from Figure 1.2. Here the data is represented using a line chart, a common chart type used to show how quantitative values change over time. In this case the items of data are represented by point marks, positioned at the intersection of the relevant x and y positions for each reporting month and channel. The attributes used here are, firstly, the connected lines that join the continuous series of values for each channel and, secondly, the distinct colours applied to distinguish each line path and associate them with their respective sales channel category.

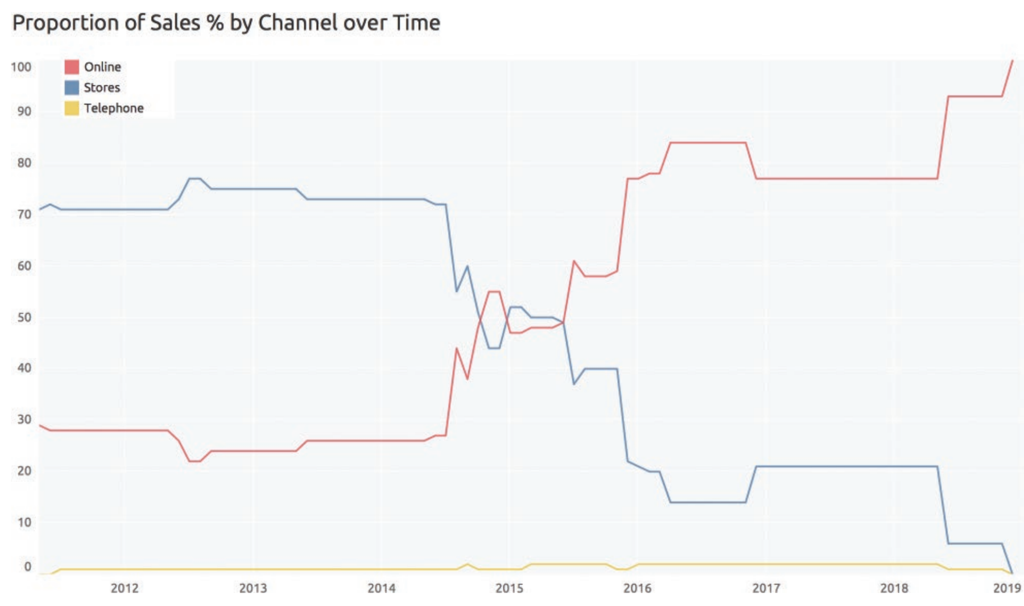


Figure 1.3 Proportion of Sales Percentage by Channel over Time

As a viewer, you scan this chart to form observations about the three sales channels individually and then compare them with each other. The comparisons made between separate channels are especially relevant for this data as the quantities shown are representative of parts of a 100% whole. This means that at any given point along the timeline, the change in value for one channel will have an effect on the values across the two others.

Consuming this data in chart form, as opposed to reading a table, enables a viewer to process clusters of multiple data points simultaneously to identify the slopes and flats, the

peaks and troughs, as well as gaps and cross-overs between lines. Though the precision of determining an individual data point (e.g. from the chart, what was the percentage share of online sales during April 2016?) is slightly diminished compared with the ease of performing the same task with data in table form, observations about the collective patterns and relationships, in turn, become more precise. The story of the rise in dominance of online sales and the related decline of store sales is immediately apparent, but what is striking here is an intense pattern of ebb and flow during the time period of mid-2014 to mid-2015, out of which the significant respective changes in trajectory of online and store sales materialised and continues.

The chart has the same data as the table, but it is represented differently. Whether this chart view is better than a table view depends on the purpose of your communication and the needs of your audience. You do not chart data because you can, you do it because it provides a window for seeing different features of data. We will explore the judgements you need to make about what you want to show your audience in Chapter 5 and, in particular, in Chapter 6 where you will learn about the wide range of established chart types that are commonly used by the visualisers of today. These charts vary in complexity and composition. Each is capable of accommodating different types of data and portraying different types of analysis. This chapter will broaden your visual vocabulary, giving you an appreciation of more ways to express your data. It will also increase the sophistication of how you go about making effective choices.

The next component of the definition is **presentation**, which concerns all the other design decisions that make up the full anatomy of any visualisation. As this text is focused on creating visualisation as a means for communicating to others, presentation concerns how we choose to 'package up' a visualisation work to impart it to an audience, irrespective of the medium or disseminating method.

Visual presentation includes design choices such as the possible application of interactivity, features of annotation, all matters around colour usage, and the composition of the work. Considering the line chart in Figure 1.3, if this was intended for the Web, you could envisage interactivity being useful to offer tooltip details of value labels as you hover over parts of each line. You could offer controls to modify the x-axis time range or filtering to hide or show different lines of interest. There are some features of annotation already on display with this chart, such as the title, the colour legend, and the x- and y-axis scales. You could also add captions to provide explanations about some of the most noticeable patterns in the data. As mentioned, colour is already used as an attribute, to distinguish the lines for each sales channel category, but the application of colour extends across every visible element including the background shading, gridline colours, and colouring of any text or labels. Finally, composition relates to the size and placement of all design elements, like the dimensions of the chart area, the alignment and size of the title, and the placement of the axis labels.

The thinking that goes into designing the full anatomy of a visualisation – combining visual *representation* and *presentation* – is inevitably interconnected. The selection of a chart type inherently triggers a need to think about the space and place it will occupy on your screen or

page; a clickable interactive feature that reveals annotated captions requires careful thought about how to style the text and what colours to use.

There are lots of seemingly small design decisions to make in visualisation, little things that add up to having a big impact. During the early stages of learning this subject it is helpful to partition your presentation thinking and tackle these design concerns as separate layers. Chapters 7–10 will explore each of these design matters separately, but in sufficient depth, profiling the options available and the factors that influence your decisions. As you gain experience and assurance, the interrelated nature of the choices you make will become more seamless and you will be stimulated by the depth of thinking demanded of you.

The final component of the definition expresses that data visualisation aims to **facilitate understanding**. Everything in this book essentially boils down to helping you accomplish this objective. We will deal with the term *facilitate* shortly, but let's focus for now on the word *understanding*.

The notion of understanding is quite broad. To best explain its relevance to data visualisation requires us, again, to turn to the perspective of a viewer.

When consuming a visualisation, a viewer will go through a process of understanding involving three phases: *perceiving*, *interpreting* and *comprehending* (Figure 1.4). These are not just synonyms for the same word, rather they convey distinctions in cognitive focus.

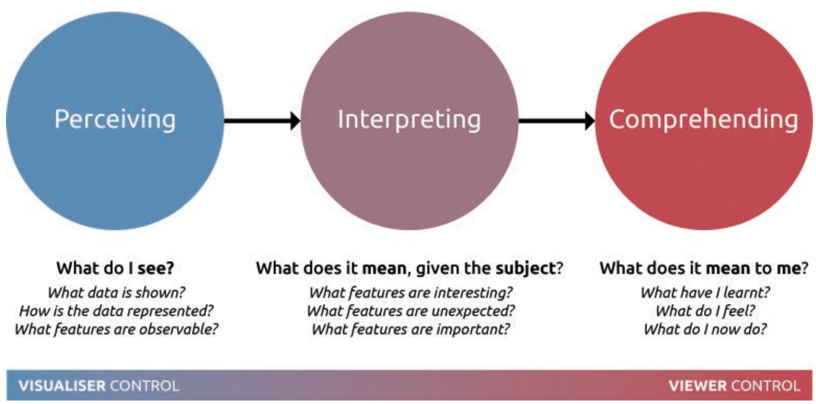


Figure 1.4 The Three Phases of Understanding

For the benefit of this illustration we will consider them to occur in a linear sequence, with successive phases being dependent on the preceding phase having been accomplished. To a viewer, consciously trying both to understand a visualisation and to extract understanding from a visualisation, these different phases will feel rather indiscernible. They might appear to occur in parallel. Viewers are human – there are occasions when rapid interpretations of a chart's headline features are made before the whole content has had a chance to be perceived first.

Let's look at the characteristics of and differences between these phases referring, initially, to an example chart (Figure 1.5) that presents some headline statistics about footballer Lionel Messi's career with FC Barcelona.

The first phase is *perceiving*, and this concerns the act of reading a chart: 'what do I see?'. A viewer decodes how the data is represented to form initial observations about the main features of the displayed data:

- What chart is being used?
- What items of data do the marks represent? What value associations do the attributes represent?
- What range of values are displayed?
- Are the data and its representation trustworthy?

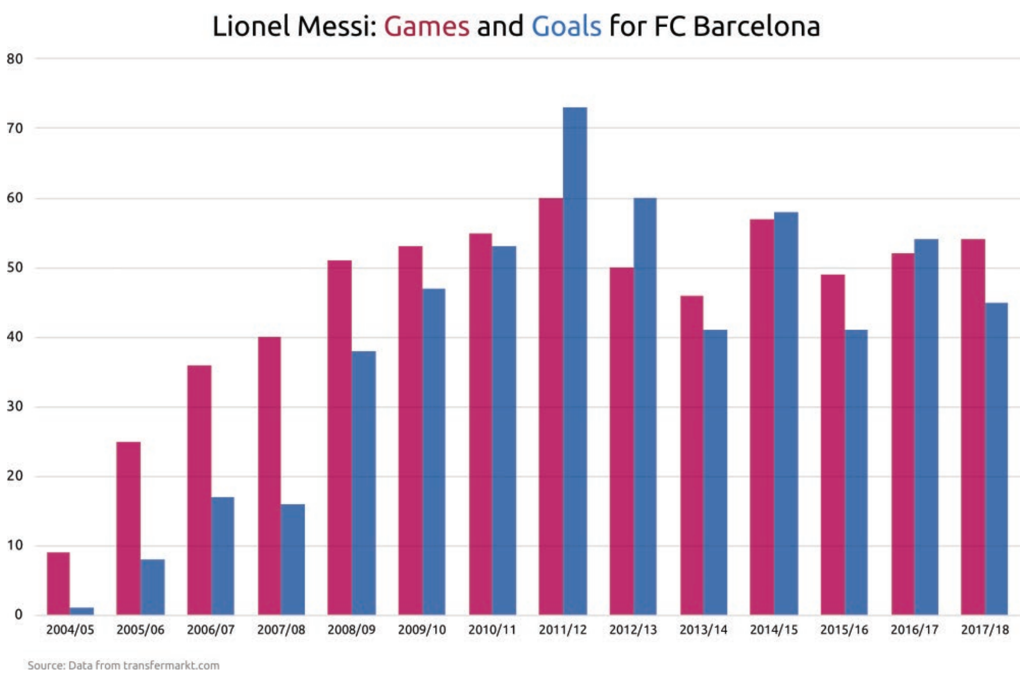


Figure 1.5 Lionel Messi: Games and Goals for FC Barcelona

Source: Data from transfermarkt.com

In the example we see a clustered bar chart showing quantitative values of pairs of categories over time. This is a chart type I am familiar with and so I feel instantly at ease with the prospect of consuming it.

I see time is plotted on the x-axis in years – or, more specifically, football seasons – and a shared quantitative measure is on the y-axis. There are two distinct categories of bars for each season, with the colour association explained by an explanation key integrated into the title. The burgundy bars show the games played in a season and the blue bars the number of goals scored. This title also helps establish clarity about what the data is showing. As the representation method is understood, initial observations begin to form about the main characteristics of the display:

- What features – shapes, patterns, differences or connections – are *observable*?
- Where are the largest, mid-sized and smallest values? (known as 'stepped magnitude' judgements).
- Where are the most and the least? Where is the average or *normal*? ('global comparison' judgements).

When scanning the chart, my eyes are drawn to the dominant bars in the middle and towards the right of the display. I am particularly interested in the highest pair of bars in 2011/12. With assistance offered by the horizontal gridlines and axis labels I can perceive with reasonable confidence that the highest number of goals scored was 73 and the most games played was 60. I can see that the burgundy bars – showing games played – are relatively stable in size since around 2008/09, but the blue bars are more erratic. The bar heights for both categories are much smaller the further left the time series goes. Looking between the categories, there is no consistency in the relationship as the burgundy bars are sometimes larger than their blue neighbours, sometimes smaller.

Interpreting, the second phase of understanding, translates these observations into quantitative and/or qualitative meaning. Interpreting involves assimilating what you have observed against what you know about the subject. What does what you have seen mean, given the subject?

- What features – shapes, patterns, differences or connections – are *interesting*?
- What features are expected or unexpected?
- What features are important given the subject?

The task of drawing interpretations from the observations I made on the chart is helped considerably by my interest in and knowledge of football. I know that if a player is scoring more than 25 goals in a season this is very good, and to score over 35 is exceptional. To achieve 50, 60 or indeed 70 goals in a season is frankly preposterous, especially at the highest level of the game. I know it is rare for a player to be scoring at a ratio of greater than one goal per game played, so the seasons where a blue bar exceeds the height of the burgundy bars represent a quite remarkable statistic. I could elaborate on some of the features I expected to (and do) see in this chart based on knowing the periods when different managers were in charge of Barcelona, which other players were in the team, and how the team performed from one season to the next. I know what to expect in terms of the classic shape of a footballer's career arc and can map that onto Messi's, anticipating that at some point – but not yet – the classic rise, peak and plateau will inevitably be followed by steady decline.

As this commentary demonstrates, a viewer's ability to perform rational interpretation will be significantly determined by factors external to the visualisation itself. The degree of knowledge viewers possess about the portrayed subject and their capacity to close a knowledge gap is fundamental. To fulfil the perceiving of a chart, viewers need the context of scale; to fulfil the interpreting of a chart, viewers need the context of subject. Furthermore, there is the matter of willingness. At the time of consuming a visualisation, not everyone has the inclination to engage with it, especially if they have no interest in a subject or if it has no immediate relevance to their needs.

My connection with the subject of football helped me understand more about the meaning of the features of data compared with other viewers who might possess no knowledge of the sport. Switching the subject from football to a completely made-up topic, but using the same chart with the same data, reinforces this. In Figure 1.6 we see a chart displaying data about the sightings of *Winglets* and *Spungles*.

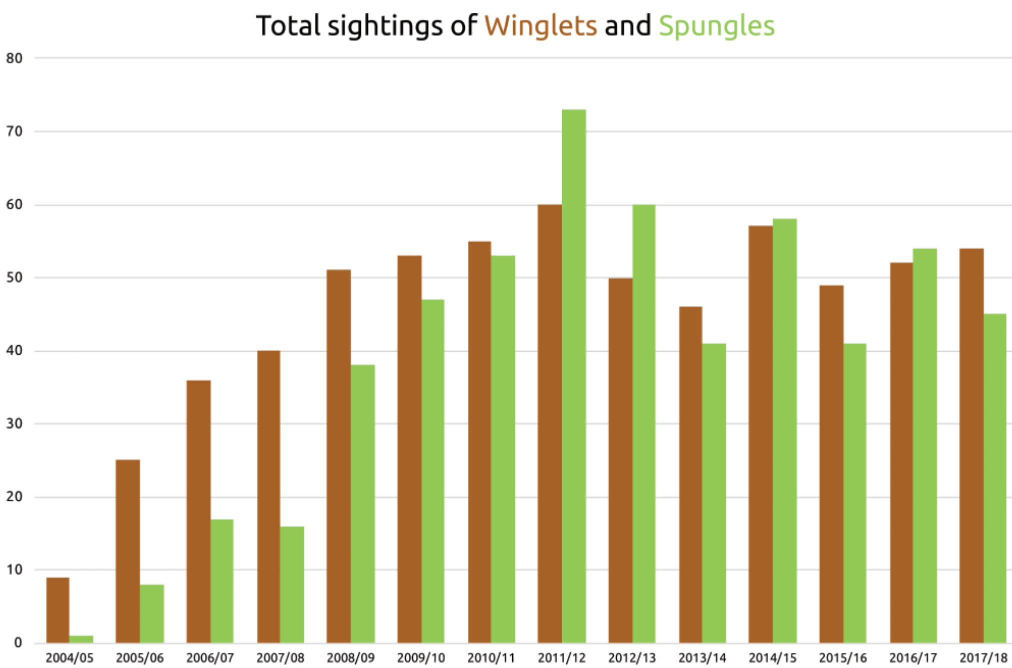


Figure 1.6 Total Sightings of Winglets and Spungles

I can still perceive the chart, observing the same features as I did when it was portraying Messi's quantities of games and goals, but as I have no knowledge of this subject I cannot interpret it. I have no idea what *Winglets* and *Spungles* are, so I cannot form any reasonable sense of what is interesting, surprising or important about the features of this display. My process of understanding stops after the perceiving phase.

As this illustrates, any deficit in a viewer's connection to a subject will fundamentally impede progress towards performing interpretation. Additionally, this may heighten the risk of the viewer drawing spurious or unsupported interpretations from a visual display.

In situations where a potential viewer might not possess sufficient knowledge of a subject, it will require the visualiser to assist in bridging the gap between observation and meaning. This can be achieved through simple design elements like the provision of captions, inclusion of headlines and astute use of colour to create emphasis, for example. The viewer must then take responsibility to learn from the assistance provided. As the purple colouring of the middle phase circle shown in Figure 1.4 denotes, forming useful and reasonable interpretations is a shared responsibility.

The final phase of understanding is *comprehending*, which is the consequence or reflective legacy of the communication experience. The viewers now consider what the interpretations mean *to themselves*. What can be inferred as being important to you about the interpretations you have made?

- What has been learnt? Has it reinforced or challenged existing knowledge? Has it been enlightened with new knowledge?
- What feelings have been stirred? Has the experience had an impact emotionally?
- What does one do with this understanding? Is it just knowledge acquired or something to inspire action, such as making a decision or motivating a change in behaviour?

In my case, the outcome of the understanding achieved from the Messi chart is nothing too dramatic or emotional. There is no direct action linked to it, rather I simply reflect on gaining a heightened impression, formed out of this data, about how sensational a footballer he has been and continues to be. For Barcelona fanatics who watch him play every week, they will have already formed this understanding. This information would only reaffirm what they already knew. To others less familiar with the subject, it might be more enlightening, but only if they had any requisite interest.

One person's 'wow' is another person's 'I knew that' is another person's 'I don't care'. Even if you have just two people in your target audience group, you have potentially two different viewer profiles. We cannot always anticipate what they do not know, what they want to know and what is the relevance to them of knowing something.

Visualising data is just an agent of communication and not a guarantor for what a viewer does with the opportunity for understanding that is presented. There are different flavours of comprehension, different consequences of understanding formed through this final phase. Many visualisations will be created with the ambition simply to inform, like the Messi graphic achieved for me, perhaps to add just an extra grain to the pile of knowledge held about a subject. Not every visualisation exists to lead a viewer towards some Hollywood-esque moment of grand discovery, surprising insight or life-changing decision. That is OK, though, as long as the outcome fits with the intended purpose, something we will discuss in more depth in Chapter 3.

Once again, the association a viewer has with the subject portrayed will greatly influence this comprehending phase. Returning to the data shown earlier about the percentage sales by channel over time for Company X, let's suppose this was a chart produced to assess the effectiveness of a corporate strategy to consolidate operations towards an online-only sales model.

The outcome of the interpretations formed from this chart might be to draw the conclusion that whatever actions were taken, they have succeeded. Depending on when the 100% online sales target was expected, it may be that this chart demonstrates complete success. It might also reveal belated success. Maybe the company was hoping for 100% online sales far sooner than when they were achieved. Conversely, the analysis shown might reveal unexpected patterns of sales. Online channels are clearly dominating, but what if the company is still maintaining the expense of running stores, with staff costs and stock tied up in what appears to be an expired model? There might be substantial costs assigned to telephone operators waiting for the phone

to ring in order to make a potential sale. But nobody is phoning, so maybe the company should look at restructuring.

All these are reasonable avenues that comprehending this data could lead to. But, at this point, I should reveal that the real context of this data actually had nothing to do with sales. That subject was picked for illustration purposes, but the data was actually about something else.

Specifically, this was data about the ever-shifting forecasts during the night of the 2016 US Election. The data values came from FiveThirtyEight, a respected website noted for its use of statistical techniques to analyse and tell stories about elections and several other data-rich subjects. The quantities relate to the 'chances of winning the presidency' forecasts for the two main party candidates, as well as a residual 'other result' percentages to make up the 100% aggregate. The temporal dimension concerned the times during the night of the election (8 November 2016), when key results were declared, influencing the changes shown in the forecasted outcome at each point.

Figure 1.7 shows the same chart as before, with the same quantities plotted and with the same design, but now reflecting the true context of the subject matter, as indicated by the updated title, colour key and x-axis scales.

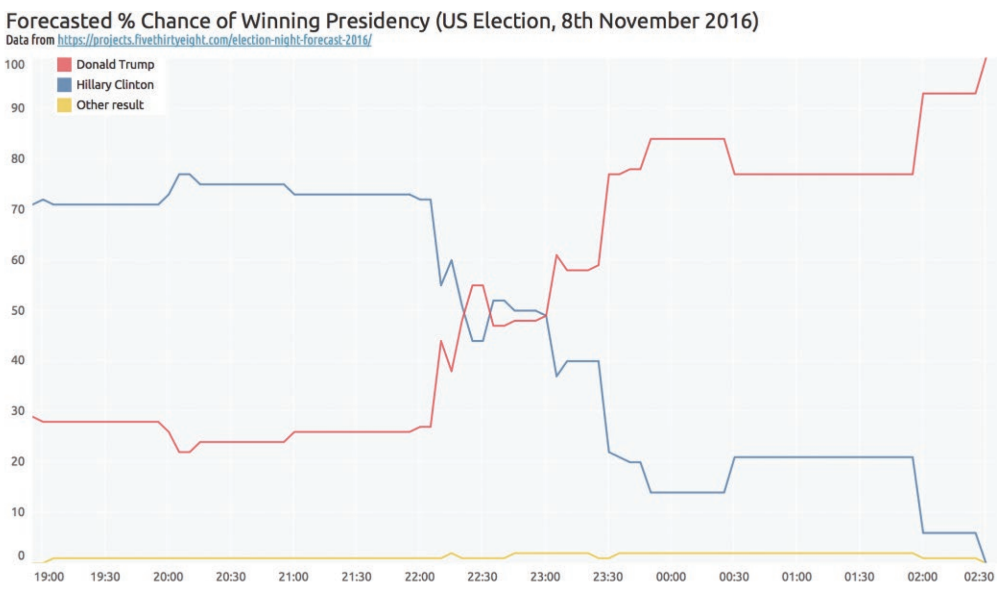


Figure 1.7 Forecasted % Chance of Winning Presidency (US Election, 8 November 2016)

Irrespective of where you sit politically, the revised context of the data portrayed in this chart will unquestionably change how you feel about what you now see. It is no longer just a routine sales chart restricted in relevance to a small group of people at Company X. It is now a visualisation about a momentous event in modern history, the outcome of which most people on the planet have some connection with or awareness of.

There are consequences of emotion to consuming this data. Some will relive the wild jubilation of their candidate's unexpected victory, others will recoil in horror at the memory of their candidate's unexpected defeat.

There are consequences of enlightenment. Some will be seeing these compelling patterns of ebb and flow for the first time, others will at least recollect this roller-coaster story playing out via TV or web coverage during the night itself.

There are also rational reactions. Consuming this chart now, many months or years later, offers the opportunity for more considered analysis, in contrast to the original setting of this data being consumed live across the USA and the rest of the world via a dynamically – and dramatically – changing forecast tracker. In the cold light of day questions can be asked (and have been) about the rigour of polling methods as well as the calculations used to create such forecasts. 'How could they be so wrong?' some have asked, while others have countered with 'How could they be expected to be more right, it's a complicated electoral system!?'

From your perspective as the visualiser, this final phase of understanding is something you will have limited control over. Everything depends. It can be frustrating for people who are learning visualisation and who just want *the* answer: 'How do I *deliver* understanding to my audience?!'

In my experience, the factors that most influence the success of a visualisation are not technical, they are contextual and, furthermore, human. Viewers are people. People are different, and people are complex. They can be irrational and unpredictable, or impassive and disengaged. You can lead a horse to water, but you cannot make it drink: you cannot force viewers to be interested in reading your work, nor to understand the meaning of what you present, nor control how they react to that experience. Even if your visualisation clearly shows action needs to be taken, you cannot guarantee the viewers will recognise there is a *need* to act, will be in a *position* to act, and indeed will know *how* to act.

It is at this point that we must recognise the ambitions and – more importantly – the limitations of what data visualisation can deliver. Returning to the definition for a final time, the illustrations we have gone through in this chapter support why the term **facilitating** is realistically the most a visualiser can do. It might feel like a rather tepid duty, something of a cop-out that abdicates responsibility for the outcome – why not aspire to achieve something more concrete than 'facilitate'?

I use *facilitate* because it gets to the heart of the tensions that visualisers face. There are times when the onus is on us, and other times when the onus is on the viewer. Visualisation design cannot change the world, it can only make it run a little smoother. Visualisers can control the output but not the outcome; at best we can expect to have only some influence on it. The rest of this book concerns how we optimise this influence.

1.2 Distinctions

Having delved into the proposed definition for data visualisation, it is now worth acknowledging some other associated terms and disciplines that you may be familiar with or aware of.

The subtleties and semantics of defining fields are recurring concerns as new technologies develop and creative techniques evolve. As participation has grown over the past decade, data visualisation has been cross-pollinated with creative and analytical sensibilities arriving from different origins. The traditional boundaries begin to blur and the practical value of preserving dogmatic distinctions reduces accordingly. Ultimately, when one is tasked with creating a visual portrayal of data, does it really matter if the creation is labelled and filed under 'data visualisation' or 'infographic' as long as it achieves the aim of helping the audience to achieve some form of understanding?

However, subject distinctions do need to be understood. It is important for people to identify with a particular discipline in which they have recognised expertise. It is therefore worth clarifying some proposed distinctions, so, once again, we are on the same page of understanding.

Infographics: The classic distinction between infographics and data visualisation concerns the format and the content. Infographics were traditionally created for print consumption, in newspapers or magazines, for example. The best infographics explain things graphically – systems, events, stories – and can often be generalised as explanation graphics. Infographics contain charts (visualisation elements) but may also include illustrations, photo-imagery, diagrams and text. These days, the art of infographic design continues to be produced for static output – as opposed to interactive – irrespective of how and where the work is published.

Earlier this decade there was an explosion in different forms of infographics. From a purist perspective, this wave of work was generally viewed as being an inferior form of infographic design. These pieces were primarily driven by marketing desire for 'clicks', above any real desire to facilitate understanding. If your motive is 'bums on seats' then I feel this is a different endeavour to pure infographics and I would question the legitimacy of attaching the term infographic to these designs; perhaps instead info-posters or tower graphics (they commonly existed with a fixed-width dimension and huge length in order to be embedded into websites and onto social media platforms) could be used. It is important not to dismiss entirely the evident – if superficial – value of this type of work, as demonstrated by the occasional viral success story. But I sense the popular interest in these forms has now waned and the authentic superior-quality infographic has managed to rise back out of this noise.

Information visualisation: Smarter people than me use labels of data visualisation and information visualisation interchangeably, without a great deal of thought for the relevant differences. The general distinction tends to be shaped by one's emphasis in focus towards either the input material (data) or the nature of the output form (information). It is common for information visualisation to be used as the term to define work that is primarily concerned with visualising abstract data structures such as trees or graphs (networks) as well as other qualitative data (therefore focusing more on relationships rather than quantities).

Information design: Information design is a design practice concerned with the presentation of information. It is often associated with the activities of data visualisation; indeed sometimes it is presented as the major field in which data visualisation belongs. Unquestionably, both share an underlying motive to facilitate understanding. However, in my view, information design has a much broader application concerned with the design of many different forms of visual communication, particularly those with an instructional or functional slant, such as way-finding devices like hospital building maps or in the design of utility bills.

Data journalism: Also known as *data-driven journalism* (DDJ), this concerns the increasingly recognised importance of having numerical, data and computer skills in the journalism field. In a sense it is an adaption of data visualisation but with unquestionably deeper roots in the responsibilities of the reporter/journalist.

Visual analytics: Some people use this term to relate to analytical-style visualisation work, such as dashboards, that serve the role of operational decision support systems or provide instruments of business intelligence. The term is also used to describe the analytical reasoning and exploration of data facilitated by interactive visual tools. This aligns with the role of exploratory data analysis that I will be discussing in Chapter 4.

Data science: As a field, data science is hard to define, so it is easier to consider it through the lens of a data scientist's duties. Data scientists are somewhat unicorn-like in that they possess – or are expected to possess – an almost preposterous repertoire of capabilities covering the gamut of demands involved with gathering, handling, analysing and presenting data. Typically, the data scientist works with data of large size and complexity. Data scientists have strong mathematical, statistical and computer science skills, not to mention astute business experience, and are also expected to possess so-called 'softer' abilities like problem solving, communication and presentation.

Scientific visualisation: This is another form of a term used by many people for different applications. Some label exploratory data analysis as *scientific visualisation* (drawing out the scientific methods for analysing and reasoning about data). Others relate it to the use of visualisation for conceiving highly complex and multivariate datasets specifically concerning matters with a scientific bent (such as the modelling functions of the brain or molecular structures).

Data art: Apart from the disputes over the merits of certain infographic work, data art is arguably the other discipline related to visualisation that has historically stirred up the most debate. Again, maybe it is reasonable to suggest the noise is quieter these days, but its sheer existence still manages to wind up certain sections of the data visualisation illuminati. Data artists work with a similar raw material in the form of data, but their goal is not driven by facilitating the kind of understanding that a data visualisation would seek. Data art is more about pursuing a form of self-expression or aesthetic exhibition using data as the paint and algorithms as the brush. As a viewer, the meaning you draw from displays of data art are entirely down to the personal interpretation it invites.

Dashboard: These are popular methods for displaying multiple visualisations and statistical information. Dashboards often take the form of some organisational instrument that offers both at-a-glance and detailed views of many different analytical and information dimensions. Dashboards are not a unique chart type themselves, but rather should be considered compositions that comprise multiple chart types.

Storytelling: This is an increasingly common term that is often misused and misunderstood, which is quite understandable. Stories are usually constructed upon some notion of movement, change or narrative. Charts showing trends or activities over a temporal plane or maps portraying spatial relationships offer displays that are most consistent with the idea of a story. A bar chart alone does not represent a story, in most people's sense of the term, but if you show a pair of bar charts to represent a before-and-after comparison, you have created a change dynamic.

Similarly, if you incorporate charts into some temporal presentation like a slideshow or video, the chart becomes a prop and a narrator may draw out the story verbally. In this case it is the setting and delivery that are consistent with the notion of storytelling, not the chart itself.

A further distinction to make is between stories that are explicitly *communicated* and stories that form through *interpretation*. The famous six-word story *For sale: baby shoes, never worn* by Ernest Hemingway is not presented as a story, rather the story is triggered in our mind when we read this passage and start to infer meaning, implication and context. A story is being presented only if it is accompanied by some explanation of the meaning of the data. Otherwise, any story derived is what the viewers form themselves.

Summary: Defining Data Visualisation

In this chapter you have been introduced to the subject of data visualisation, learning a definition that will shape much of the structure and content of this book:

> *The visual representation and presentation of data to facilitate understanding.*

The different components that form this definition have been explained, with particular focus on the nuances around facilitating understanding. The three distinct phases of understanding were described:

- **Perceiving:** what do I *see*?
- **Interpreting:** what does it *mean*, given the subject?
- **Comprehending:** what does it *mean* to me?

The second section explained some of the distinctions and overlaps with other related disciplines, supplementing the glossary provided in the Introduction.

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

2

The Visualisation Design Process

In this second chapter I will outline the data visualisation design process around which the book's chapters are arranged. You will learn why using a process approach is important to organise and optimise your thinking – taking you from the initial spark of curiosity, through wrangling with data, to juggling the myriad options that shape a design solution.

The process organises the activities into a sequence of manageable chunks so that the right things are tackled in the right order. You cannot expect just to land on a great solution by chance if your working practices are chaotic and confused. You will be aided by some additional practical tips and good habits to employ across the whole process.

The quality of your decision making is the main difference between a visualisation that succeeds and one that fails. To maximise the effectiveness of facilitating understanding for your audience, the sectional parts of the chapter will introduce the three principles of good visualisation design.

2.1 Design Process: Organising Your Decision Making

For those new to the field, one of the first things to grasp is the idea that any notion of *perfect* in data visualisation does not exist. It can prove simultaneously frustrating and liberating to learn that there are good and bad solutions, but there are no perfect ones. To have perfect you need immaculate conditions that are free of pressure, constraint or flaw. That is how things operate now in real life. There will always be demands pushing and pulling you in different directions. There will be shortcomings in the data that frustrate you or limitations in technical ability that impede you. As described in Chapter 1, people, as recipients, introduce a diversity of need that realistically cannot always be fulfilled. Recognising that perfect is unobtainable helps unburden us from a nagging sense that somehow we might have missed finding *the* perfect solution. There will never be just one single possible solution to a problem.

The central premise in this book is that decision making is the key competency in data visualisation: namely, effective decisions, efficiently made. To accomplish this you need to follow a design process that organises your thinking and is underpinned by robust principles to optimise your thinking.

We will discuss principles shortly, but firstly let's look briefly at the design process overall (Figure 2.1).



Figure 2.1 The Four Stages of the Data Visualisation Design Process

Across the four stages that make up this process there are two main phases. The first three stages, presented in Part B of this book through Chapters 3 to 5, involve activities that I describe as concerning the 'hidden thinking' of data visualisation. These stages cover the preparatory work that informs *what* you are visualising, for *whom* and, crucially, *why*:

1. Formulating your brief: planning, defining and initiating your project.
2. Working with data: gathering, handling and preparing your data.
3. Establishing your editorial thinking: defining what you will show your audience.

The second main phase of the process sits entirely with stage 4 and this involves developing your design solution, the visual manifestation of the preparatory work you have conducted. This stage is concerned with the *how*.

The five distinct design layers that make up the anatomy of any visualisation solution – data representation, interactivity, annotation, colour and composition – are covered in Part C of this book, in Chapters 6 to 10 respectively. As explained earlier, a detailed treatment of technical activities is beyond the scope of this text.

I am not going to describe these process stages in more depth here – the next eight chapters exist to do that. Instead, here are some observations about why it is important to follow a design process.

Reducing the randomness of your approach: The value of this design process is that it shapes your entry and closing points. How do you start a process? How do you know when you have finished? As I have mentioned, the sheer extent of things you will have to think about, even with simple projects, can be quite an overwhelming prospect. This approach breaks down key stages into a connected system of thinking that will help progress your work and preserve cohesion between your activities. It incrementally leads you towards developing a solution, with each stage building on the previous one and informing the next.

Every project is different: Every visualisation presents new challenges. Even if you are just re-producing the same report every month, no two instances of that report will involve the exact same context. Just by having one extra month of data, for example, may expose you to larger values, smaller values, new values and expired values. Whether you have simple data, or vast

amounts of complex data, two hours or two months, the process you follow will always be the same. You should follow the same sequence of thinking regardless of the size, speed and complexity of your challenge. The main difference is that any extremes in the circumstances you face will amplify the stresses at each stage of the process and place greater demands on the need for thorough, effective and timely decision making.

> 'I tend to keep referring back to the original brief (even if it's a brief I've made myself) to keep checking that the concepts I'm creating tick all the right boxes. Or sometimes I get excited about an idea but if I talk about it to friends and it's hard to describe effectively then I know that the concept isn't clear enough. Sometimes just sleeping on it is all it takes to separate the good from the bad! Having an established workflow is important to me, as it helps me cover all the bases of a project and feel confident that my concept has a sound logic.' **Stefanie Posavec, Information Designer**

Adaptability: The term *process* contrasts considerably with *procedure*. The process outlined in this book provides a framework for thinking, rather than instructions to learn and follow. A good process should offer adaptability and remove the inflexibility of a defined procedure. In any visualisation project, you will need to respond to revised requirements, additional data that emerges, or a shift in creative direction. A good process safeguards adaptability and cushions the impact of changing circumstances like these. Although the activities presented in this book are in a linear arrangement, there will always need to be room for iteration. There will be plenty of occasions when you have to revisit decisions or redo activities in a different way, especially if you make mistakes. What is more important in these situations is how gracefully you fail and how quickly you recover.

Protect experimentation: The process approach I am advocating is not overly systematic and does not compromise on allowing space for experimentation. When there are pressures on time, the need to focus and avoid distraction is understandable. Aspiring to reduce wasted effort and improve efficiency is entirely reasonable, but one must still seek out opportunities – in the right circumstances – for imagination to blossom. In reality, few projects will offer too much scope for far-reaching creative exploration, but when an opportunity presents itself for you to work on a subject that befits creativity, you should embrace it. And do not forget to enjoy it!

The first occasion, not the last: Each activity you commence across the distinct stages in the process will likely represent the *first* occasion you pay attention to these matters, but not the *final* occasion. Think of the sequencing as being akin to a trickle-down effect. Take, for instance, the recurring concern about thinking about your audience. You will first encounter the need to define a profile of your anticipated audience's characteristics during the first stage of the process, 'Formulating your brief'. However,

> 'I truly feel that experimentation (even for the sake of experimentation) is important, and I would strongly encourage it. There are infinite possibilities in diagramming and visual communication, so we have much to explore yet. I think a good rule of thumb is to never allow your design or implementation to obscure the reader understanding the central point of your piece. However, I'd even be willing to forsake this, at times, to allow for innovation and experimentation. It ends up moving us all forward, in some way or another.' **Kennedy Elliott, Graphics Editor, *National Geographic***

the concern about what they know, what they need to know, and how interested they will be will reoccur right through to the end. Concerns like these should never drop off your radar. The list of concerns will only build, but the intention is that the process gives you the best chance of keeping all the necessary plates spinning for as long as they need to be.

Across the book there are frequent vignettes of advice and useful tips for you to adopt to get the most out of working through this process. These are informed by interviews with people working in the field, as well as from my own practical experiences, and are provided with each topic in the book. There are some recommended habits that are applicable to all stages in this process, relevant to novices or experienced visualisers alike, as follows.

Time management: Any creative work quickly swallows up all the available time. You get tempted to try things, to explore different ideas, to attempt one final pass at seeking out interesting features of your data. It is easy to be consumed by the stretching demands of the activities across this process. As you then reach a deadline you either sink or swim: for some the pressure of the clock ticking is crippling, especially impacting their creative thinking; others thrive on the adrenaline it stirs, sharpening their focus as a result. Regardless of how you respond to looming deadlines, good planning is vital.

Time management is the essence of good planning. It keeps a process cohesive and on track. From experience working on different projects your ability to anticipate how much time to allocate to different activities will improve. That said, each project introduces its own profile of demands, so always find time before you set off to estimate where your likely commitments will be most required. Do not forget to factor-in time for easily neglected responsibilities, such as supervisor meetings, Skype calls, research and file management.

Mindsets: Irrespective of the type of visualisation you are working on, your process will involve a mixture of conceptual and practical activities. Sometimes these will be allocated across a team, exploiting the range of talents at different times through the process. On other occasions you will be working alone, and the diversity of these activities will stretch your mind considerably. Sometimes you are thinking, sometimes you are creating; sometimes you need to be creative, sometimes you need to have an eye for detail.

> 'You need a design eye to design, and a non-designer eye to feel what you designed. As Paul Klee said, "See with one eye, feel with the other."' **Oliver Reichenstein, Founder of Information Architects (iA)**

- *Thinking:* The duties here will be conceptual in nature, requiring imagination and judgement, such as formulating your curiosity, defining your audience's needs, reasoning your editorial perspectives, and making decisions about viable design choices.
- *Doing:* These are active duties that engage the brain through more practical undertakings, such as sketching ideas, conducting research, holding discussions with a client, or checking data.
- *Making:* These are more hands-on constructive duties characterised by using tools for activities like handling data, creating charts, and designing presentation features.

For the scope of this book, the focus is largely on thinking. I find the notion of brain 'states' relevant here, especially the 'alpha' state. This is the state our mind is in, most commonly,

when we feel especially relaxed. Occupying this state helps heighten your imagination and thought process. I find I do some of my most astute thinking in the shower or just before going to sleep at night. These are the occasions when I am most likely drifting into a relaxed state. I find the same conditions when undertaking long train journeys or flights. I use it to help contemplate the progress I am making on a task. It lets me escape the noise present when doing more practical tasks.

Documenting: It is mawkish to claim the humble pen and paper are the most important tools for visualisers. After all, unless you are producing artisan hand-drawn work, technical applications will be more applicable for most of your process. However, pen and paper will prove to be a real ally to help you document thoughts and capture sketches. Do not rely on your memory; if you have a great idea, sketch it down. You do not need great artistry, you just need to get things out of your head and onto paper, particularly if you are collaborating with others. If you are fortunate to be fluent with a tool and find it more natural to use that for 'sketching' ideas than pen and paper, then this is absolutely fine, as long as it is the quickest medium to do so.

Whether using pen and paper, or a tool like Word or Google Docs, note-taking is a useful habit to develop. It helps you document important details such as:

- task lists with details of deadlines and precedents;
- information about the sources of data you are using;
- details of complicated calculations or manipulations you have applied to your data;
- a log of any assumptions you have made;
- terminology, abbreviations, acronyms – technical properties of your data that are crucial to its understanding;
- questions and answers you have received or are yet to;
- issues or problems you have experienced or can foresee;
- wish lists of features or ideas you would like to explore;
- sources of inspiration, like websites or magazines you discover;
- ideas you have had or rejected.

Note-taking is more easily preached about than done. I am the least competent of note-takers, but I have found a way to make it a forced habit and it does prove valuable.

Communication: Communication is a two-way activity. It is about listening to stakeholders and to your audience: what do they want, what do they expect, what ideas do they have? In particular, what knowledge do they have about your subject? Communication is about speaking to others: presenting ideas, updating on progress, seeking feedback, sharing your thoughts about possible solutions, and

> 'Because I speak the language of data, I can talk pretty efficiently with the experts who made it. It doesn't take them long, even if the subject is new to me, for them to tell me any important caveats or trends. I also think that's because I approach that conversation as a journalist, where I'm mostly there to listen. I find if you listen, people talk. (It sounds so obvious, but it is so important.) I find if you ask an insightful question, something that makes them say "oh, that's a good point," the whole conversation opens up. Now you're both on the same side, trying to get this great data to the public in an understandable way.' **Katie Peek, Visualisation Designer and Science Journalist**

promoting and selling your work (regardless of the setting, you will need to do this). You cannot avoid the demands of communicating, so do not hide behind your laptop – get out there and interact with people who can help or whom you can help.

Associated with the need for good communication skills is the importance of research. You cannot know everything about your subject, about the meaning of your data, about the relevant and irrelevant qualities it possesses. As you will see later, data itself can only tell us so much; often it just tells us where interesting things might exist, not what actually explains why they are interesting. Talk to smart people who know a subject better than you or people who do not know the subject but are just smart.

> 'Research is key. Data, without interpretation, is just a jumble of words and numbers – out of context and devoid of meaning. If done well, research not only provides a solid foundation upon which to build your graphic/visualisation, but also acts as a source of inspiration and a guidebook for creativity. A good researcher must be a team player with the ability to think critically, analytically, and creatively. They should be a proactive problem solver, identifying potential pitfalls and providing various roadmaps for overcoming them. In short, their inclusion should amplify, not restrain, the talents of others.' **Amanda Hobbs, Researcher and Visual Content Editor**

Attention to detail: The process you follow embodies the concept of the 'aggregation of marginal gains'. You need to sweat the small stuff. Even if many of your decisions seem tiny and inconsequential, they deserve your full attention. Like note-taking, the importance of checking every detail may not be a natural trait for some. However, errors found in your work can be damaging and will certainly undermine your audience's trust, as you will learn about shortly. I know through experience how one mistake can undermine the integrity of an entire project, even if this feels unfair and disproportionate considering everything that was correct. Start every project with a commitment to eliminate mistakes and learn from the pain when you fail. It is not easy: I have no doubt I will leave at least one mistake in this book and it will haunt me. It can help, if you are so immersed in your own work and become *blind* to it, to seek others to help you.

'Kill your darlings': A recurring consequence of facing so many decisions in visualisation is the need to demonstrate the discipline of *not* doing something. It is easy to applaud oneself over brilliant ideas, but occasionally these ideas you have invested in deeply just will not work out. Even though you have invested heavily in time and emotional energy, do not be stubborn. When something is not working, learn to kill it. Otherwise, such preciousness will impede the quality of your work. Being blind to things that are not working, or ignoring constructive feedback from others, will prove destructive.

Learn: Reflective learning is about looking back over your work, examining the output and evaluating your approach. What did you do well? What would you do differently? How well did you manage your time? Did you make the best decisions you could, given the constraints that existed? Learn from others. Read how other people undertake their visualisation challenges. Maybe share your own? You will find you truly learn about something when you find the space to write about it and explain it to others. Write up your projects, present your work to others and, in doing so, this will force you to think 'why did I do what I did?'

Also use reflective learning to find *your* process. What is presented in this book is proposed, not imposed. If you cannot get this approach to fit your personality, your project's purpose, or the rhythm of how you need to work with others, modify it. We are all different. Take this as a recommended framework but then bend it, stretch it and make it work for you. As you become more experienced (and confident through having been exposed to different challenges) the activities involved in data visualisation design will become second nature. You will probably become blissfully unaware of even observing a process.

2.2 Design Principles: Optimising Your Decision Making

If the goal of data visualisation, as defined in the first chapter, is to *facilitate understanding*, all judgements made through the design process have to contribute to accomplishing this.

> 'I say begin by learning about data visualisation's "black and whites", the rules, then start looking for the greys. It really then becomes quite a personal journey of developing your conviction.' **Jorge Camoes, Data Visualisation Consultant**

Most choices are relatively clear cut and basing your judgement on common sense, informed by the first three preparatory stages, will be entirely reasonable. However, for more nuanced situations, when there might be several complex options presenting themselves, you will face a dilemma. Making a choice will need more than just common sense. This is when it helps to consult a framework of design principles.

In data visualisation there are relatively few universal rules to follow. There are evidence-based, useful suggestions that nudge you towards 'always do this' and 'never do that', but even they are exposed to legitimate breaking point. This is because each decision that needs to be made is accompanied by many contextual dependencies.

The principles that inform my own visualisation design convictions originated from beyond the boundaries of this subject. Dieter Rams was a German industrial and product designer most famously associated with the Braun company. Around the late 1970s and early 1980s, he was becoming concerned about the state and direction of design thinking and, given his prominent role in the industry, felt a responsibility to challenge himself, his own work and his own thinking. He posed the simple question: 'Is my design *good* design?' By dissecting his work in response to this question he conceived ten principles that symbolised the important aspects of what he considered to be good design (Figure 2.2).

1. Good design is innovative.
2. Good design makes a product useful.
3. Good design is aesthetic.
4. Good design makes a product understandable.
5. Good design is unobtrusive.
6. Good design is honest.
7. Good design is long-lasting.
8. Good design is thorough down to the last detail.
9. Good design is environmentally friendly.
10. Good design is as little design as possible.

Figure 2.2 Dieter Rams' 'Ten Principles of Good Design'

In 'De architectura', a thesis on architecture written around 15 BC by Marcus Vitruvius Pollio, a Roman architect, the author declares that the essence of quality in architecture is framed by the social relevance of the work, not the eventual form or workmanship towards that form. He states that good architecture can only be measured according to the value it brings to the people who use it. In a 1624 translation of the work, Sir Henry Wotton offers a paraphrased version of one of Vitruvius' most enduring notions that a 'well building hath three conditions: firmness, commodity, and delight'. An updated translation of this would read as 'sturdy, useful and beautiful'.

Collectively, these separate sources have informed the three key principles shown in Figure 2.3 that I believe apply to any judgement of effectiveness in data visualisation.



Figure 2.3 The Three Principles of Good Visualisation Design

As you go through the process, these principles will guide the choices you make. Let's look at them in more detail, framing them in relation to Rams' ten principles of good design as well as Vitruvius' three.

Principle 1: Good Visualisation Design Is Trustworthy

This first principle is to make your visualisation trustworthy, which maps directly onto one of Dieter Rams' ten principles (Figure 2.4) and embodies Vitruvius' desire for *sturdy*. This is about reliability and is achieved by securing and sustaining the trust of your audience.

Figure 2.4 Mapping 'Trustworthiness' onto Rams' Ten Principles



This is presented as the first of the three principles because it is about the fundamental legitimacy of a data visualisation. Without trust, the opportunity to facilitate understanding vanishes. You can disregard any value from achieving 'accessible' and 'elegant' design; if you lose trust, you lose your audience. Game over.

There is an important distinction to make about the relationship between *trust* and *truth*. Achieving trust is an aim, presenting truth is an obligation. There should be no compromise here. You should never create work you know to be misleading, through either its content or its representation. You should never claim something presents the truth if it cannot be reasonably supported. The difference between a truth and an untruth should be beyond dispute. The fact that it is not, these days, is a sad indictment of modern society. Nevertheless, the imperative for truthfulness must be clear.

> 'Good design is honest. It does not make a product appear more innovative, powerful or valuable than it really is. It does not attempt to manipulate the consumer with promises that cannot be kept.' **Dieter Rams**

The difficulty for a visualiser comes when there are potentially multiple different, but legitimate, versions of a 'truth' within the same data or subject context. This muddies things somewhat. A glass that is half full is also half empty. Both viewpoints are objectively truthful, but the one you choose to focus on is subjective. In data visualisation there is rarely a single view of the truth. You make the call about what is most relevant in the context of your work. This is something we will explore in Chapter 5 about editorial thinking.

Even though the choice made might be impartial, it is the act of choosing that creates an unavoidable form of bias. When you choose to do one thing you are usually choosing not to do something else. Deciding to show the forecasts for a political party winning an election changing over time using a line chart might also incorporate a decision *not* to show how the forecast looks geographically, assuming the data was available. Any visualisation will be an aggregation of decisions among which many are shaped by reasonable subjectivity. No visualisation is purely objective, even if it seems to portray this quality.

Rather than get consumed by the inevitability of biases rippling through your work, and perhaps seeing this as good reason *not* to undertake it, your focus is better directed towards ensuring your chosen path is *trustworthy*. In the absence of a single objective truth, what can you do to secure trust in your subjectively selected truth?

Trust must be earned, but it is hard to secure and easy to lose. As the translation of a Dutch proverb states, 'trust arrives on foot and leaves on horseback'. Trust is something a visualiser must try to nurture through accuracy and transparency, eliminating doubts or legitimate dispute from a viewer. Easier said than done, though, as visualisers have only a certain amount of control over this because our audience is people. Yes, them again. A visualisation can be truthful but viewed, unreasonably, as being untrustworthy. Conversely, a visualisation that might not be truthful might still be trusted (perhaps a more dangerous outcome that opens a separate discussion, but one for other books to tackle). Neither of these are satisfactory experiences: the latter we can and should avoid; the former is something we strive to overturn.

Let's consider an example that illustrates the fragility of trust. In Figure 2.5, the chart shown plots the number of murders committed using firearms in Florida over a given period of time. The data is framed around the enactment of the 'Stand Your Ground' law in 2005. The chart uses an inverted vertical y-axis with a red-filled area occupying the space beneath the x-axis baseline, growing downwards as the number of deaths increases. Some of the peak values are at 1990 and 2007.

When this piece was published, many commentators hastily cried foul, remarking on how the inversion of the y-axis had deceived them. They had mistaken the red area as the background and saw the data as formed by the 'white mountain' emerging in the foreground. In misreading the chart, they were instead seeing peak values as being those for 1999 and 2005, the highest points of this apparent white mountain. This illusion is caused

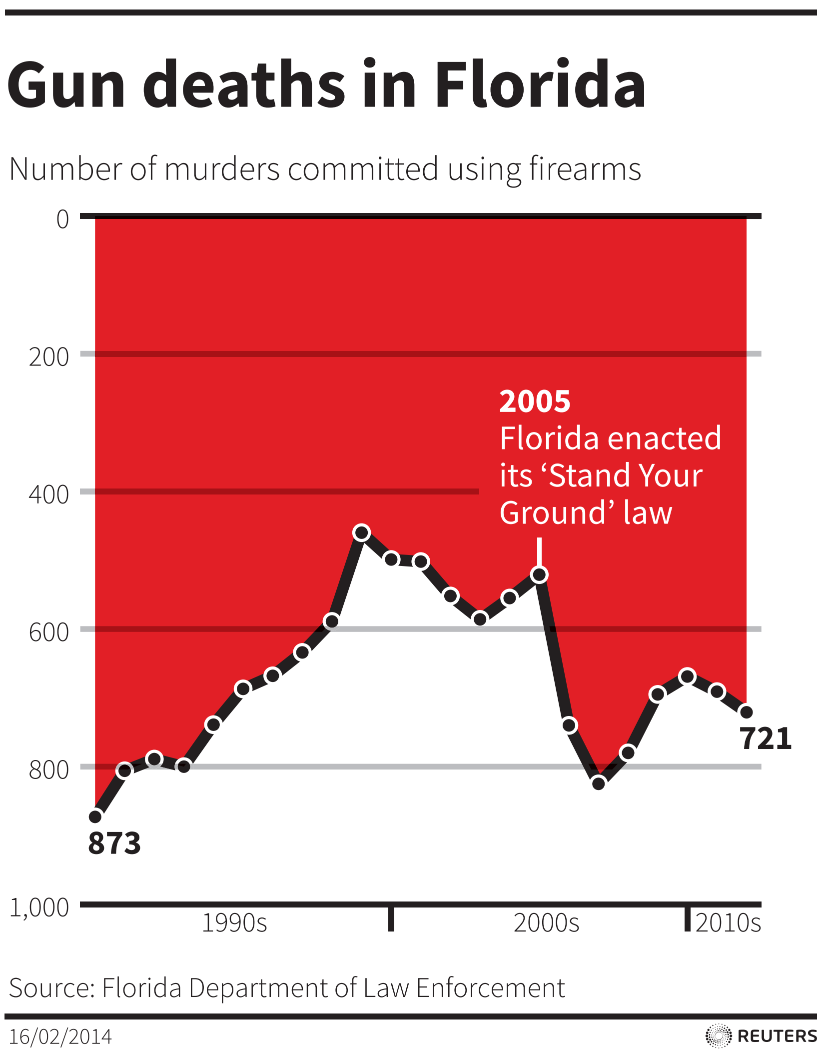


Figure 2.5 Gun Deaths in Florida (Reuters Graphics)

by an effect known as *figure–ground* perception whereby a background form (the white area) can become inadvertently recognised as the foreground form, and vice versa with the red area seen as the background. Despite eventually reading the chart and being able to discover the correct view of the chart, for many viewers, any trust had been lost: they felt they had been tricked. An accusation exacerbated, no doubt, by the emotive nature of the subject: any chart about gun crimes will stir passions regardless of its form.

Although the approach to inverting the y-axis may not be entirely conventional, it was a legitimate approach. The problem was arguably caused by the visually prominent gridline for 1000 which inadvertently framed the 'white mountain' but, creatively speaking, attempting to convey the effect of dribbling blood was a plausible metaphor to pursue.

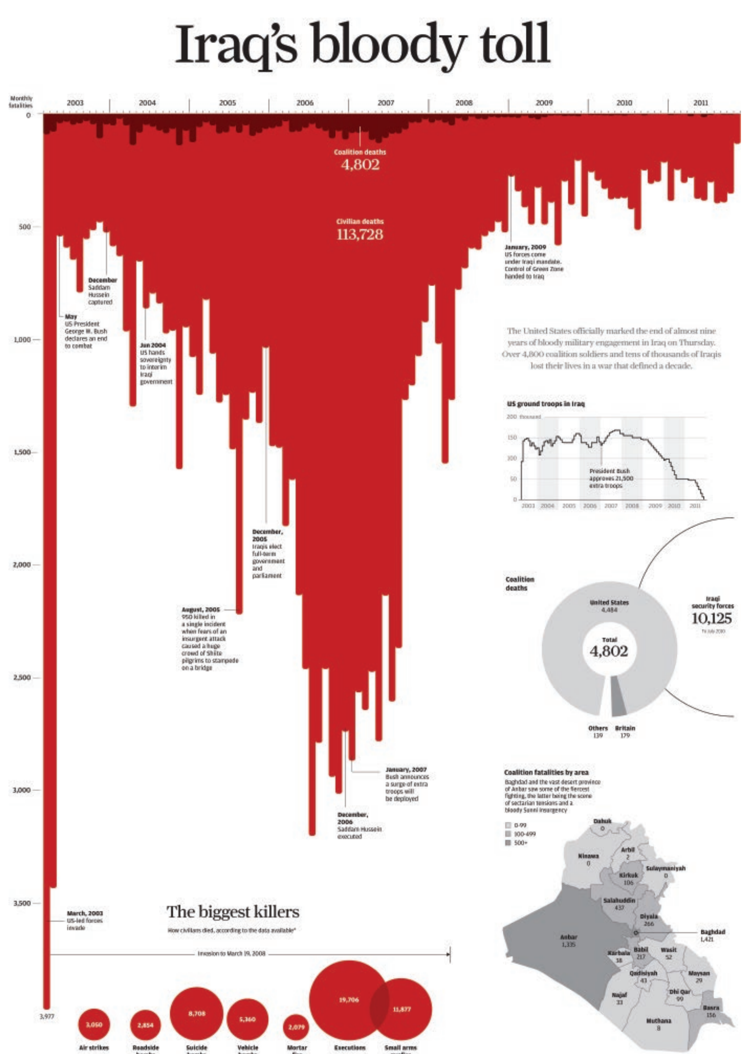


Figure 2.6 Iraq's Bloody Toll, by Simon Scarr (*South China Morning Post*)

It was emulating a prominent and celebrated visualisation work, from several years ago, showing the death toll during the Iraq conflict (Figure 2.6). Being inspired and influenced by the techniques demonstrated by other visualisers is something to be encouraged as an important way of furthering our skills.

The key point is that there was no intention to mislead. The lack of trust expressed by some was the consequence of a well-intended set of design decisions. It demonstrates how vulnerable trust is, especially in the pressured environment of a newsroom where the output of work is relentless and there is often only a single opportunity to publish a given piece to a huge, widespread audience. Even in this era of largely digital media platforms, it is hard to intercept, withdraw and revise work. 'You don't get a second chance to make a first impression', as the saying goes.

Is the Handling of the Data Reasonable and Faithful to the Subject?

> 'Data and data sets are not objective; they are creations of human design. Hidden biases in both the collection and analysis stages present considerable risks [in terms of inference].' **Kate Crawford, Principal Researcher at Microsoft Research NYC**

Trustworthiness is a cause that should guide all your decisions, not just those that emerge in the design-focused final stage of the process. The principles are framed as design principles because it is through your design work that all your decisions will visually materialise. Earning trust is something that reaches right the way back to the earliest preparatory task.

It is during the first stage that the initial seeds are sewn for how you might creatively handle your subject matter. As you have just witnessed, if you are working on potentially emotive topics, this will only heighten the potential exposure to prejudgement and opinion. As you will learn in Chapter 3, when considering your subject matter and establishing your ideas about the purpose of your work, there will be some contexts that lend themselves to exploiting the emotive qualities of your subject but others that will not. Trust will be jeopardised if you have misjudged the tone of voice.

Working with data, the second stage of the process, is arguably where trust is most at stake. You are the custodian with a responsibility for being faithful to the data you have and the subject it embodies. You need to be careful with your handling of the data and transparent with what you decide to do with it. There are critical questions you may need to answer to ensure your approach to handling the data is reasonable, such as the following:

- How was the data collected: from where and using what method?
- What calculations or modifications have you applied?
- Have you made significant assumptions or applied any specific counting rules?
- What criteria were used for the data values you decided to include and exclude from the display?

Does the Representation and Presentation Design Have Integrity?

A fundamental tenet of data visualisation is never to deceive the receiver. Avoiding possible misunderstandings, inaccuracies, confusions and distortions is of primary concern when thinking about the integrity of how your work is both represented and presented. There are many possible features of visualisation design that can lead to varying degrees of mistrust, whether intended or not, as exhibited in the previous example; I will explore many of these further later in the book:

- Sometimes charts are used in ways that effectively corrupt their usage. An example would be using pie charts to display percentage parts that exceed 100%.
- Unreliable functional experiences with interactive projects. Does the solution work and, specifically, does it work in the way it promises or is expected to do (such as the speed of loading)?
- Missing annotations like clear titles, introductions, axis titles, labels, footnotes and data sources. All these features help a viewer understand what he or she is consuming; when they are missing it can lead to confusion and suspicion.
- Mistakes made with any statistics or captions presented will tarnish the perceived accuracy of the entire work.
- The quantitative axis of a bar chart should not be 'truncated'. That is, the baseline origin value should be zero, otherwise the resulting display will distort the perceived bar size judgements.
- The size of geometric areas, such as circle sizes, can sometimes be miscalculated by using diameter as the basis of size variation rather than shape area. This results in the quantitative values being disproportionately sized.
- When a chart based on two dimensions of data is presented in 3D form but consumed in 2D format (such as a static display on a screen or in print), this decorative design choice distorts the perceived value sizes – you cannot adjust your viewpoint to accommodate perspective, process distance, or see obstructed features. Thus 3D should only be considered when there are dynamic means for a viewer to change his or her viewpoint in order to navigate around a 3D form and see it from bespoke 2D perspectives.
- The aspect ratio (height vs width) of a line chart's display can affect the perceived steepness of connecting lines which reveal the trends of a continuous series of values over time. If the chart area is too narrow the steepness will be embellished; too wide and the steepness is dampened.
- When portraying spatial analysis through a thematic map, there are different mapping projections which translate a spherical globe into a flattened 2D map. The mathematical treatment applied to this translation can significantly alter the perceived size or shape of regions, potentially distorting their perception. More on this in Chapter 10.
- The size or sequencing in the layout of a work might raise suspicions if seemingly important contents are diminished in the visual hierarchy, such as pushed to the bottom or shrunk in size.

The examples in Figures 2.7 and 2.8 display two charts showing the same analysis but presented differently. Take a moment to reflect on how much trust you feel is achieved by these respective pieces. For context, both were extracted from articles discussing issues about home ownership, so they would normally be presented alongside written analysis in their original published form.

Figure 2.7 Housing and Home Ownership in the UK (ONS Digital Content Team)

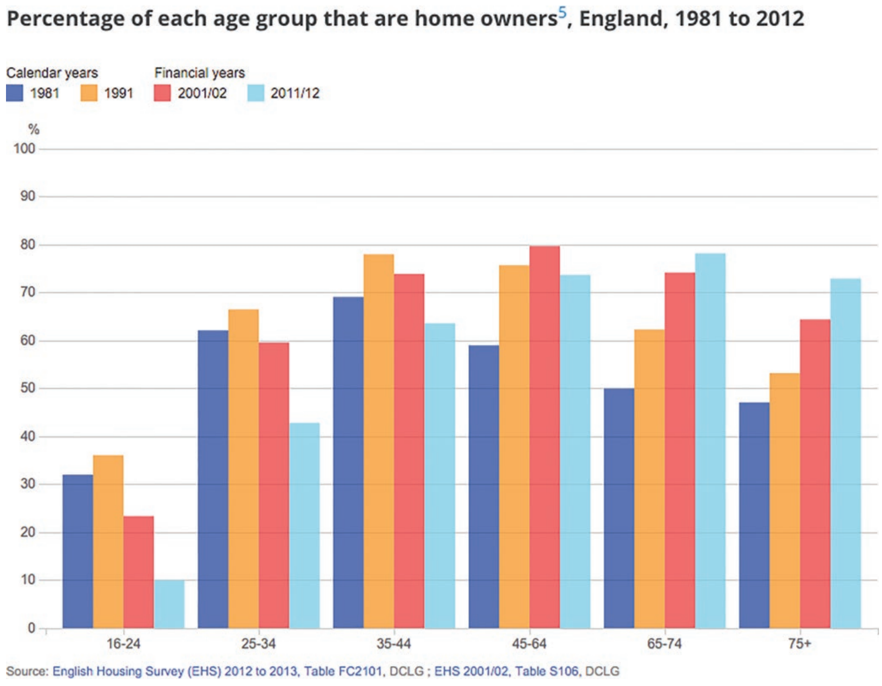
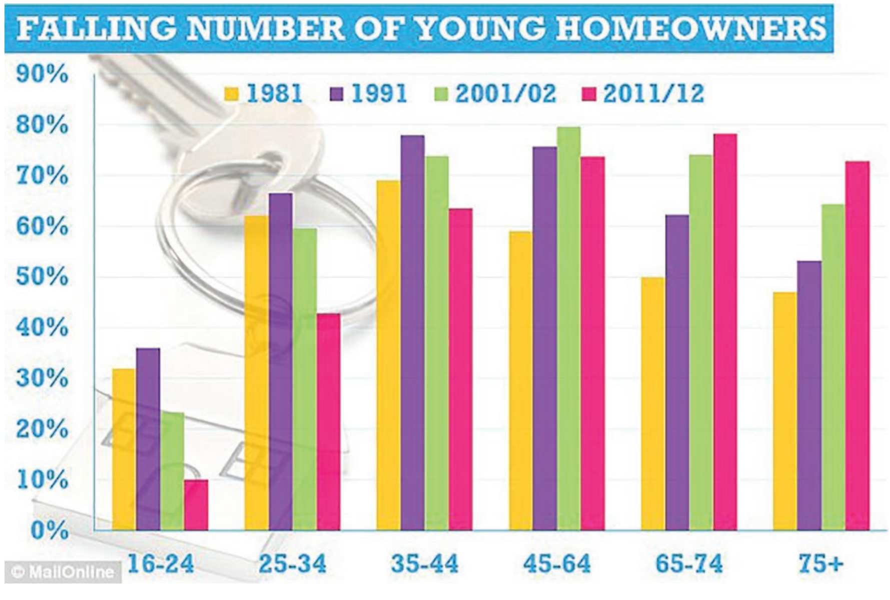


Figure 2.8 Falling Number of Young Homeowners (*Daily Mail*)



As I have said, both charts use the same data and use the same chart type to represent the analysis; they even arrive at the same summary finding. However, in my view, the first chart, produced by the UK Office for National Statistics (ONS), commands greater credibility and authority, and therefore far more of my trust, than the second visualisation, produced by the *Daily Mail*.

The first reason for this opinion concerns how the ONS piece is more informative and transparent about the data and subject. Whereas the *Daily Mail* piece refers to the ONS as the source of the data, it fails to include any more details about the data source, information which is included on the ONS graphic alongside other features like the subtitle, an explanation about the yearly periods. colour choices for the bars. The option to see and download the associated data is unique to the ONS chart as it was published on the Web, so it is unfair to contrast the absence of this feature on the *Daily Mail* graphic.

The second reason is more instinctive and influenced by my personal taste. The colours used in the ONS graphic are reserved but aesthetically engaging and convey a certain assuredness. By contrast, the *Daily Mail*'s colour palette feels needy. It seems to be craving attention with sickly sweet coloured sticks. The house key image in the background is visually harmless but feels lazy and derivative. The typeface, font size and colouring of the text feel cheap. The ONS text feels polite and conveys authoritativeness. These presentational features are stylistic and therefore more weighted towards being matters of pursuing 'elegance' in your design, as we will describe shortly, but the three principles are unquestionably interconnected in places.

Overall, my opinion about the level of trust I hold for these pieces is partially reasonable and partially irrational. Crucially, it is also greatly influenced by the prejudices I bring to the encounter. I do not trust the *Daily Mail* as a source of any information, whereas I do trust the ONS. It is hard to undo those feelings and biases that we bring to the viewing process. The platform and location in which your work is published (e.g. website or source location) will prove influential: visualisations encountered in already-distrusted media will create obstacles that are hard to overcome.

Principle 2: Good Visualisation Design Is Accessible

This second principle is to make your visualisation accessible, which maps onto three of Dieter Rams' ten principles of good design (Figure 2.9) as well as Vitruvius' desire for *useful*.

1. Good design is innovative.	6. Good design is honest.
2. Good design makes a product useful.	7. Good design is long-lasting.
3. Good design is aesthetic.	8. Good design is thorough down to the last detail.
4. Good design makes a product understandable.	9. Good design is environmentally friendly.
5. Good design is unobtrusive.	10. Good design is as little design as possible.

Figure 2.9 Mapping 'Accessibility' onto Rams' Ten Principles

Accessibility in visualisation design is concerned with giving your audience access to useful understanding. It must be relevant to the subject and relevant to their needs. This needs to be achieved in a way that does not require undue effort to perceive, interpret and comprehend.

Is the Portrayal of the Data and the Subject Relevant?

The first aspect of accessibility concerns the relevance of the visualisation you are portraying to your audience. Judging relevance is a subjective and contextually driven matter relating to the potential usefulness of your visualisation: am I providing my audience with access to the most useful understanding about this subject? Relevance is a somewhat shifting concept that is, in part, based on qualities such as interestingness and pertinence.

It is also, fundamentally, shaped by what you *actually* have available to present to your audience. You might create the most beautifully designed visualisation, but if nobody finds the specific analysis you choose to portray about a subject relevant, any motivation your audience had to engage with your work may be undermined.

> 'The key difference I think in producing data visualisation/infographics in the service of journalism versus other contexts (like art) is that there is always an underlying, ultimate goal: to be useful. Not just beautiful or efficient – although something can (and should!) be all of those things. But journalism presents a certain set of constraints. A journalist has to always ask the question: How can I make this more useful? How can what I am creating help someone, teach someone, show someone something new?' **Lena Groeger, Science Journalist, Designer and Developer at ProPublica**

Imagine you are visiting a city for the first time and you ask a passer-by for help with directions to the main railway station. Unfortunately, they cannot guide you to the station, but they do know how to get to the main library. In these circumstances, what was *useful* for you to learn was not matched by what was possible for the passer-by to impart. Directions to the library do not give you access to the understanding you actually need and is therefore irrelevant. However, suppose you later discover that the railway station is across the road from the library; the information available about how to get to the library is now instantly promoted to being relevant. It is just a shame it is too late.

Any judgement of relevance will also be determined by the level of content sophistication. In the Introduction's glossary of terms, you will have seen the distinction between terms like *complex* and *simple*. As a visualiser, you might misjudge the level of sophistication required by your audience and oversimplify a complex subject. Maybe there were lots of interesting perspectives about the subject that you could and should have included, but instead you just presented one simplified chart and that might disguise many of the key nuances about the subject. Equally, maybe you made a great effort to include multiple related chart views that provide a wide coverage of different aspects of your subject, when what your audience need is just a simple chart and some quick headline insights. In each case you have failed to provide suitable access to the relevant content.

This is evidently a hard thing to judge, especially when you have a varied audience with diverse interests. You can only do so much, and you certainly should not expect to get it right for every potential viewer. But it is a crucial matter to care about. Most of the topics and activities covered in Chapters 3, 4 and 5 are concerned with guiding you towards a reasonable judgement of relevance.

Is the Representation and Presentation Design Suitably Understandable?

In contrast to relevance, the suitability of design looks at usefulness from a different perspective. This is less about the consequence of a visualisation's use and more about understanding how to use it.

Accessibility in design is fulfilled by removing any design- and content-related obstructions faced by your viewers. Expressed another way, thinking of the opposite of accessible (confusing), can you ensure a viewer avoids a confusing experience with your work? Confusion is the friction between the *act* of understanding (effort) and the *achieving* of understanding (reward).

What constitutes *minimum friction* is shaped, inevitably, by context and the characteristics of the audience, which makes the notion of accessibility a somewhat variable concept. It is not always possible to eliminate friction, which is why a judgement of 'suitable' friction is a pragmatic perspective to take. The efforts need to feel proportional to the rewards on offer. Not every process of understanding can or should be quick and simple, but to achieve accessibility means to eliminate unnecessary delays to this process.

Demonstrating empathy for your audience, appreciating the setting in which they encounter your work and how they need to use it are at the heart of accessible design thinking. Here are some of the most crucial factors.

> 'We should pay as much attention to understanding the project's goal in relation to its audience. This involves understanding principles of perception and cognition in addition to other relevant factors, such as culture and education levels, for example. More importantly, it means carefully matching the tasks in the representation to our audience's needs, expectations, expertise, etc. Visualizations are human-centred projects, in that they are not universal and will not be effective for all humans uniformly. As producers of visualizations, whether devised for data exploration or communication of information, we need to take into careful consideration those on the other side of the equation, and who will face the challenges of decoding our representations.' **Isabel Meirelles, Professor, OCAD University (Toronto)**

Understanding a subject: What your audience know and do not know about a subject will have a significant bearing on the degree to which they consider a visualisation to be accessible. What is considered inaccessible to one audience group could be fully accessible to another. Reading a street sign written in Japanese is entirely inaccessible to non-Japanese speakers, yet it would be fully accessible to someone who knows the language. If that street sign is encountered in Tokyo it is contextually appropriate, but in Newcastle upon Tyne it would not be.

As I demonstrated in the first chapter when interpreting the charts about football and 'Winglets and Spungles', if you do not understand a subject, this instantly raises the chances of confusion through ignorance. Interpretation is prevented.

Though existing knowledge is important, the principal property of subject matter that most influences accessibility is the intellectual level it embodies. In other words, is it a complicated, complex or simple subject for anyone to grasp? This leads us again into a discussion about the semantics of language, but these differences are crucial in how we make judgements about the suitability of accessibility:

- *Complicated* relates to subject knowledge or a skill that is typically intricately technical, probably unique and difficult to understand. It requires a certain level of intellect or inherent talent to do so. The mathematics that underpinned the Moon landings is complicated. The inner workings of a boiler are complicated. Making Baked Alaska (successfully) is complicated. The knowledge or skill in question is acquirable and the learning involved surmountable, if steep, but only achieved through lots of time, hard work and, usually, with assistance from external expertise.
- *Complex* is associated with systems or contexts that have no perfect conclusion or even no end state. Managing relationships is complex. Same with parenting: there are books and people offering advice but there is no rulebook for how to do it well all the time – no definitive way of accomplishing it. The elements of parenting might not be necessarily complicated – such as remembering to cut the crusts off Sergio's sandwiches to avoid a tantrum – but the interrelated natures of events and pressures are shaping and colliding to make it feel very hard to master.
- *Simple*, for the purpose of this book, concerns a matter that is inherently easy to understand. It may be small in dimension and scope, meaning there is not a lot of knowledge to acquire and it is unlikely to require lots of practice to sustain, irrespective of prior experience. It is also quite isolated in that it does not have other interconnections affecting its state.

> 'Strive for clarity, not simplicity. It's easy to "dumb something down," but extremely difficult to provide clarity while maintaining complexity. I hate the word "simplify." In many ways, as a researcher, it is the bane of my existence. I much prefer "explain," "clarify," or "synthesize." If you take the complexity out of a topic, you degrade its existence and malign its importance. Words are not your enemy. Complex thoughts are not your enemy. Confusion is. Don't confuse your audience. Don't talk down to them, don't mislead them, and certainly don't lie to them.' **Amanda Hobbs, Researcher and Visual Content Editor**

When working with a complex or complicated subject, your instinct might be to seek to simplify it. Simplifying is a reductive process that translates a complex or complicated state into a simplified form, usually by eliminating details or nuance. There are situations that will warrant making the process of understanding quicker and easier, though this is not a universal goal.

Not everything can or should be simple. The process of simplification might risk the subject being oversimplified to the point of obscurity. In removing important subtleties and technicalities this can be just as detrimental to the perceived accessibility as

leaving a complex or complicated subject too intellectually demanding. What if your audience are sufficiently sophisticated with the capability and motivation to handle the learning process required in grasping a hard topic? By simplifying things, they would be denied that learning opportunity and denied access to relevant understanding. An audience in this case may justifiably feel patronised when faced with an oversimplified portrayal.

When considering the level of your subject matter and the nature of your analysis, if you do not think your audience will understand what you are presenting, you have a choice: to simplify or clarify.

- Simplify when your audience do not have the knowledge or capacity to handle a complicated subject and do not need to acquire deep understanding about it.
- Clarify when your audience do not have the knowledge but do have the capacity to handle a complicated subject, with assistance. Provide features of annotations to explain what the subject is about, how to read it, what features are significant, and what it all means. Do not underestimate the capacity of your audience to be willing and able to grasp topics they have no prior understanding about. Often, this is what they want out of a visualisation experience.

A further risk to creating confusion is through carelessness and complacency: do not assume domain expertise among all your audience with the use of ambiguous acronyms, abbreviations or technical language. Explain what needs to be explained. Include annotations for titles, scales and units, explaining data sources and colour associations. These are all features that contribute a great deal towards eliminating confusion.

Representation design: As well as subject complexity, another issue to consider is representation complexity. That is, the perceived complexity of uncommon or unfamiliar chart types. Not every chart we encounter is familiar. And when something is not familiar, it is understandable how this exposes a risk of confusion.

Sometimes, the lack of familiarity is a trigger to blame the visualiser: 'Why did they use a chart I've never seen before?' This deficit in knowing how to read a new or unfamiliar chart type is not a failing on the part of the viewer, it is simply the viewer's lack of prior exposure to these different methods. But what if there was good reason to use that chart? It might be the most astute way to portray the most relevant analysis about a subject. Sure, it might be visually complicated, but this might be the only way to show it.

Everything is new once. This is why we learn and why we are capable of learning. To overcome chart types that are unfamiliar a viewer must be provided with the means to learn how to perceive and interpret a chart. This prospect can naturally frustrate some people who see it as a critical obstacle they are unwilling to entertain. Even with the best intent and the provision of helpful guidance, if a viewer is simply unwilling to make the effort, you have little further influence in overcoming this blockage.

Time: This concerns the characteristics of the encounter and the duration of time or attention viewers might have available to process understanding. You need to consider if, at the point of consuming a visualisation, the viewers are in a pressured situation. Are they in a rush? Do they

need quick insights or is there scope for a more prolonged engagement? Maybe they do not have time to read about the background to a subject or learn how to read a given chart because there are direct operational decisions at stake. If so, only through the immediacy of understanding will this visualisation be considered effective. Furthermore, if you create an interactive solution with an excessive number of features, the richness in functionality may undermine its usage. The audience may lack the necessary desire to make an effort to interrogate and manipulate the display. More clicks can equate to more obstacles towards understanding.

Attitude and emotion: As viewers, on occasion we are not in the right mood. We might be tired and lazy, or we have had a particularly bad day. At these times the prospect of engaging with even the most compelling and well-designed visualisation might feel too much. I spend all my days looking at visualisations and can recognise how indifferent I feel towards work when fatigue kicks in.

An extension of mood is confidence. Sometimes the audience may not feel sufficiently equipped to embark on a visualisation if it is about an unknown subject, even if assistance is available. It might be somewhat irrational, but they do not wish to engage, especially if it takes them beyond their comfort zone in terms of the demands asked of them to interpret and comprehend.

Presentation design: In what format will your viewers need to consume your work? Are they going to need work created for a print output or a digital platform? Does this need to be compatible with a small display as on a smartphone or a tablet? Will any viewers have visual impairments that need accommodating?

If what you create is consumed away from its intended native format, like viewing a large infographic with small text on a mobile phone, that will impede the experience for the viewer. How and where your work are consumed will often be beyond your control and you cannot mitigate for every eventuality.

Principle 3: Good Visualisation Design Is Elegant

The third principle is to make your visualisation elegant, which maps again onto three more of Dieter Rams' ten principles of good design (Figure 2.10) and Vitruvius' desire for *beautiful*. Elegance is concerned with creating an aesthetic that will appeal to your audience and endure, sustaining positive sentiment throughout the experience, far beyond just the initial moments of engagement.

Figure 2.10 Mapping 'Elegance' onto Rams' Ten Principles

1. Good design is innovative.	6. Good design is honest.
2. Good design makes a product useful.	7. Good design is long-lasting.
3. Good design is aesthetic.	8. Good design is thorough down to the last detail.
4. Good design makes a product understandable.	9. Good design is environmentally friendly.
5. Good design is unobtrusive.	10. Good design is as little design as possible.

Elegant design is presented as the third principle for good reason. Any choices you make towards achieving 'elegance' must not undermine the accomplishment of trustworthiness and accessibility in your design. Indeed, the pursuit of the other principles often already leads to a certain elegance as a by-product.

> 'When working on a problem, I never think about beauty. I think only how to solve the problem. But when I have finished, if the solution is not beautiful, I know it is wrong.' **Richard Buckminster Fuller, Celebrated Inventor and Visionary**

Is the Representation and Presentation Design Appealing?

The pursuit of elegance is as elusive as a practical definition. What gives something an elegant quality? The adjectives that surface in my mind are *stylish*, *dignified* and *graceful*. Elegance has a timelessness that transcends more fleeting notions like *fancy* or *cool*.

Elegance is most conspicuous when it is missing. This is when a visualisation's design lacks cohesion and inspiration, especially across the colour and composition elements that so inform its appearance. By contrast, as expressed by Rams' principle 'Good design is as little design as possible', elegant design accelerates you to the content and to understanding.

In his book *The Shape of Design*, designer Frank Chimero references a Shaker proverb: 'Do not make something unless it is both necessary and useful; but if it is both, do not hesitate to make it beautiful.' In serving the principles of trustworthy and accessible design, you will hopefully have covered both the *necessary* and *useful*. As Chimero suggests, if we have served the mind, our heart is telling us that now is the time to think about beauty. There are several components of design thinking I believe contribute to achieving elegance in design.

Eliminate the arbitrary: As with any creative work, good editing is a hugely valuable skill. Every single design decision you make – every dot, every pixel – should be justifiable. Nothing that remains in your work should be considered arbitrary, based on random tastes, nor redundant, offering superfluous value. These will distract and, worse, may distort the process of understanding. Even if your choices are not based on empirical reasoning, you should still be able to offer justification for every feature that is included as well as any significant feature excluded.

> '"Everything must have a reason." A principle that I learned as a graphic designer that still applies to data visualisation. In essence, everything needs to be rationalised and have a logic to why it's in the design/visualisation, or it's out.' **Stefanie Posavec, Information Designer**

Eliminating the arbitrary should not be confused with the pursuit of minimalism, which is a brutal approach that strips away the arbitrary and then cuts deeper. In the context of visualisation, minimalism can be an unnecessarily savage and austere act that may be inappropriate with the style of work needed.

Thoroughness: A dedicated visualiser should be prepared to agonise over the smallest details and want to resolve even the smallest pixel-width inaccuracies. The desire to treat your work with this level of attention demonstrates respect for your audience: you want them to be able to experience quality, so pride yourself on precision. Not all decisions share

the same significance, but we need to attend to every single decision equally, caring about the small details. Do not neglect to check things and do not cut corners by not testing. It will be worth it. We are, though, only human and not every single issue can always be detected and eradicated.

Style: This is another hard concept to pin down, especially as the word holds different meaning to different people. It has been somewhat tarnished by the age-old complaints around something demonstrating style over substance. When it feels like style over substance has been at the heart of decision making, the consequence will usually prove to be an obstructed or distorted experience. Developing a style is a manifestation of elegant design. The decisions around colour selection, typography and composition are all matters that determine your style. So too does experimentation in the deployment of different representation techniques or interactive features. The development of a style creates a degree of consistency and reliability in your strongest design values that can be repeatedly deployed. It is something that needs time to develop as you find your design voice.

> 'You don't get there [beauty] with cosmetics, you get there by taking care of the details, by polishing and refining what you have. This is ultimately a matter of trained taste, or what German speakers call fingerspitzengefühl ("finger-tip-feeling")'. **Oliver Reichenstein, Founder of Information Architects (iA)**

> 'I suppose one could say our work has a certain signature. *Style*, to me, has a negative connotation of "slapped on" to prettify something without much meaning. We don't make it our goal to have a recognisable (visual) signature, instead to create work that truly matters and is unique. Pretty much all our projects are bespoke and have a different end result. That is one of the reasons why we are more concerned with working according to values and principles that transcend individual projects and I believe that is what makes our work recognisable.' **Thomas Clever, Co-founder of CLEVER°FRANKE, a Data-Driven Experiences Studio**

Many news and media organisations actively seek to devise their own style guides to help visualisers, graphics editors and developers navigate through the choppy waters of design thinking. This is a conscious attempt to foster *consistency* in approach as well as create *efficiency*. In these organisations, the pressure of tight timescales from the perpetual demands of the news cycle means that creating efficiency is of enormous value. In removing the burden of having always to think from scratch about their design choices, the visualisers are left to spend more time on the fundamental challenge of *what* to show and not get bogged down by *how* to show it. The most effective styles stand out as instantly recognisable: there is a reason why you can instantly pick out the work of the *New York Times*, the *Guardian* or the *Financial Times*.

Decoration should be additive, not negative: The decorative arts are historically considered to be an intersection of that which is useful and beauty. The term **decoration** when applied to data carries a different connotation. It is often used in criticism of the perceived dressing up of data using superfluous visual flourishes to provoke attention, usually when content is uninteresting or wafer thin in its substance.

In moderation, visual embellishments can offer effective means for securing and sustaining the appeal of an audience. We will consider the role of ideas in Chapter 3 where the headline advice is always to be primarily led by data and your audience, not your ideas. However, there are occasions in the design process when you should embrace creative flair, novelty and fun. People like nice things. Sometimes viewers crave something that stirs a more upbeat and upfront emotional engagement. A singularity of style is a dull existence for all of us.

In certain circumstances you may need to consider employing aesthetic seduction to create an appeal form that attracts viewers and encourages them to engage with a subject they might not otherwise have found relevant. This could involve the use of novel visual or functional devices that attract and perform a useful role.

This is especially the case when your ideas are congruous with the subject matter or key message. The works featured in Figures 2.11 and 2.12 show appealing enhancements to the background presentation and representation styling. The choices in each case are harmonious with the respective subjects of cocoa prices ('The KitKat' bar chart) and online razor sales (bar charts created by scraping away lengths of shaving foam). In each case these design choices offer quite a charming design solution that does not undermine the message. They supplement the information presented without obstruction or distraction.

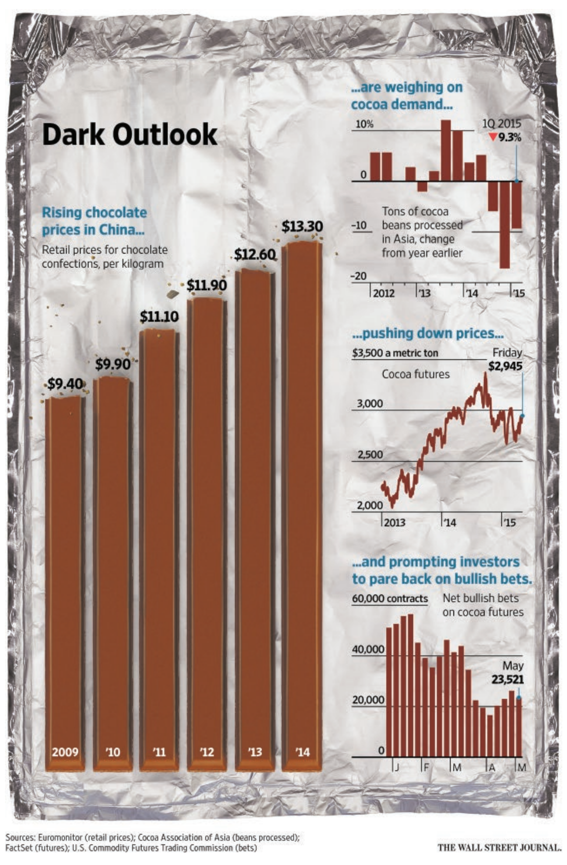


Figure 2.11 Asia Loses its Sweet Tooth for Chocolate, by Graphics Department (*Wall Street Journal*)

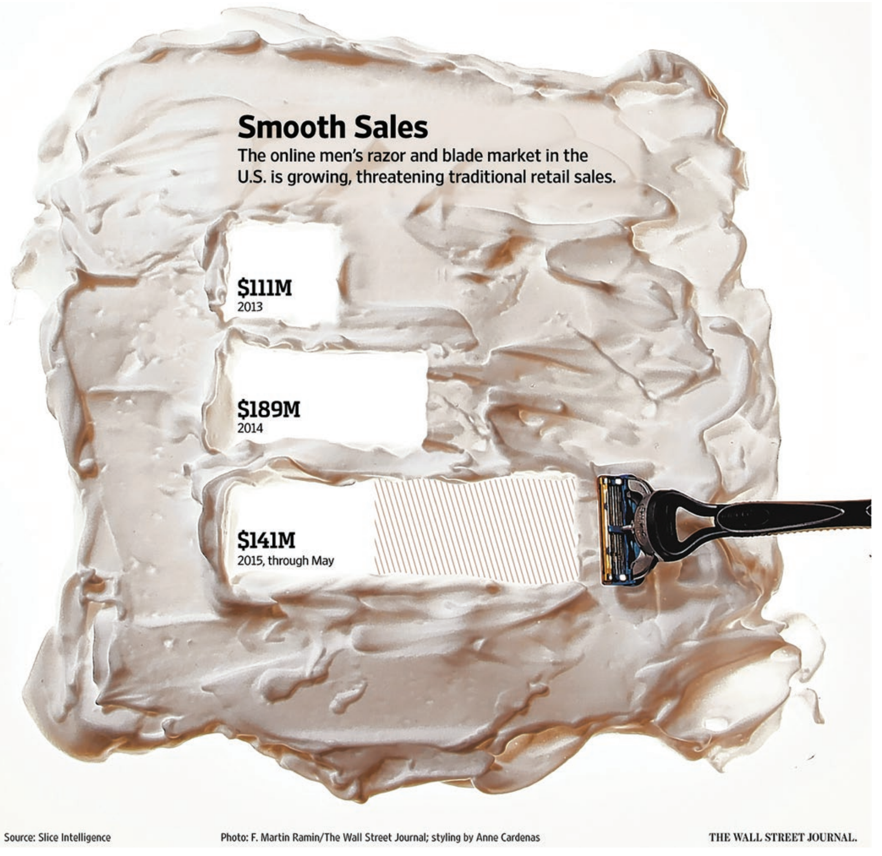


Figure 2.12 Razor Sales Move Online, Away from Gillette, by Graphics Department (*Wall Street Journal*)

'I love the idea of Edward Tufte's assertion that "Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space." But I found that when I developed magazine graphics according to that philosophy, they were most often met with a yawn. The reality is that *Scientific American* isn't required reading. We need to engage readers, as well as inform them. I try to do that in an elegant, and refined, and smart manner. To that end, I avoid illustrative details that distort the core concept. But I'm happy to include them if the topic could benefit from a welcoming gesture.' **Jen Christiansen, Graphics Editor at *Scientific American***

Some may argue that viewers will be encouraged to engage with a visualisation if it is relevant to them and they should not need to be seduced by novel appearance choices. Indeed, if they need to be convinced to look at something, maybe they should not be considered to be the intended audience? Perhaps in a business or operational setting, the needs of individuals, roles and groups are much more clear cut. Elsewhere, in the real world, there are usually more nuanced considerations about how best to connect to a potentially diverse audience demographic. Indeed, as a viewer, your interest in a subject may only materialise as a consequence of experiencing a visualisation. It might not have existed beforehand as a motivating prerequisite, and without some initial sense

of attraction or intrigue felt towards the prospect of a visualisation, a viewer may miss the chance to discover this connection.

To conclude this discussion about principles, let me explain why I feel three of Rams' original ten do not quite fit as *universal* ideals for data visualisation (Figure 2.13).

1. Good design is innovative.
2. Good design makes a product useful.
3. Good design is aesthetic.
4. Good design makes a product understandable.
5. Good design is unobtrusive.
6. Good design is honest.
7. Good design is long-lasting.
8. Good design is thorough down to the last detail.
9. Good design is environmentally friendly.
10. Good design is as little design as possible.

Figure 2.13 Considering Rams' Non-universal Principles

Good design is innovative: Most visualisations use tried and tested charting methods that have been in play for years. Unlike in product design, for example, it is not necessary for visualisers to conceive constantly new forms of representation or techniques for presentation. You might have a personal desire to be innovative, aligned to personal goals about the development of your skills, perhaps through rethinking how to tackle previous projects. It is not that data visualisation is never about innovation, just that it is not a universal principle.

That said, there are of course circumstances when innovation is important. In the context of limitations or imposed constraints, in order to overcome a particular challenge, innovation materialises. You also need it when established solutions fail to resolve new problems.

I sometimes try to look at restrictions in a positive light because of this. Consider the circumstances faced by Director Steven Spielberg while filming *Jaws*. The early attempts to create a convincing-looking shark model proved to be so flawed that for much of the film's scheduled production Spielberg was left without a visible shark to work with. Such were the diminishing time resources that he could not afford to wait for a solution to film the action sequences, so he had to work with a combination of props and visual devices. Objects being disrupted, like floating barrels or buoys and, famously, a mock shark fin piercing the surface, were just some of the tactics he used to create the suggestion of a shark rather than actually show it. Eventually, a viable shark model was developed to serve the latter scenes but, as we all now know, in not being able to show the shark for most of the film, the suspense was immeasurably heightened. This made it one of the most enduring films of its generation. The necessary innovation that emerged from the limited resources and increasing pressure led to a solution that surely transcended any other outcome that would have emerged had there been freedom from restrictions. Embrace circumstances that heighten your need to be innovative; just do not feel it is a mandatory pursuit for all visualisation contexts.

Good design is long-lasting: The translation of this principle to the context of data visualisation can be taken in different ways. 'Long-lasting' could be related to the desire to

> 'I'm always the fool looking at the sky who falls off the cliff. In other words, I tend to seize on ideas because I'm excited about them without thinking through the consequences of the amount of work they will entail. I find tight deadlines energizing. Answering the question of "what is the graphic trying to do?" is always helpful. At minimum the work I create needs to speak to this. Innovation doesn't have to be a wholesale out-of-the box approach. Iterating on a previous idea, moving it forward, is innovation.' **Sarah Slobin, Visual Journalist**

preserve the ongoing functionality of a digital project, for example. It is quite demoralising how often browser bookmarks pointing to old visualisations have now elapsed or how often digital works expire due to a lack of sustained support. The evolution of technology also risks rendering older functionality obsolete. Thankfully, the long history of print visualisation and infographic work in particular has a legacy that is simpler to preserve, in many respects.

Another way to interpret 'long-lasting' is in the durability of the technique. Bar charts or line charts, for example, are always useful, always being used, always there when you need them. 'Long-lasting' can also relate to the rejection of current fashions, preserving a timeless approach to design thinking that chimes with the discussion about elegance.

I feel 'long-lasting' most closely applies to the subject matter and the data portrayed in a visualisation. Expiry in the accuracy of data about an activity that has since changed undermines a project's long-lasting potential. This is particularly the case with subjects concerning current affairs. Analysis about the loss of life during the Second World War is timeless because nothing is now going to change the nature or extent of the underlying data (unless new discoveries emerge). Analysis of the highest grossing movies today will change as soon as new big movies are released and time elapses. So, again, the idea of long-lasting is context specific, rather than being a universal goal for data visualisation.

Good design is environmentally friendly: This is of course a noble aim, but the relevance of this principle has to be positioned again at the contextual level, based on the specific circumstances of a given project. If your work is to be printed, the resources used will undermine that project's environmental friendliness. Developing a rich interactive that is being constantly hammered by users around the world places a burden on the hosting server and the associated energy supply. Specific judgements about the scope of environmental impact of visualisation work realistically resides with the protagonists and stakeholders involved.

Finally, a comment about the need for a visualisation to be *memorable*. This is often proposed as a universal aim in data visualisation, but I disagree. If the seamless accessibility of a visualisation leads to its also being memorable, then wonderful. If the elegance of your design thinking, possibly including certain memorable visual flourishes, leave such a legacy in the mind of the viewer, then this will be a terrific by-product of your work. As an objective in itself, achieving memorability has to be considered, again, at the contextual level based on the specific goals of a given piece of work and taking into consideration the capacity of the intended audience.

Summary: The Visualisation Design Process

Design Process

In this chapter you were introduced to the design process, the sequence of activities around which the book's contents are organised:

1. Formulating your brief: planning, defining and initiating your project.
2. Working with data: gathering, handling and preparing your data.
3. Establishing your editorial thinking: defining what you will show your audience.
4. Developing your design solution: making design choices about how you represent and present what it is you want to show your audience.

It explained why a process is important to follow:

- It reduces the randomness of your approach.
- It offers adaptability to accommodate changing requirements and circumstances.
- It protects the value of experimentation.
- Each stage reached represents the first occasion you will start to undertake that activity, not the last.

Design Principles

Where the process offers efficiency, design principles ensure effectiveness. The second section introduced three key principles to help build the clarity of your convictions around the difference between effective and ineffective visualisation design:

1. Good data visualisation is *trustworthy*: Is it reliable? Is the portrayal of the data and the subject faithful? Do the representation and presentation design have integrity?
2. Good data visualisation is *accessible*: Is it usable? Is the portrayal of the data and the subject relevant? Is the representation and presentation design suitably understandable?
3. Good data visualisation is *elegant*: Is it aesthetic? Is the representation and presentation design appealing?

General Tips and Tactics

You were also presented with some general tips ahead of putting the process into practice:

- The importance of good time management.
- The need to occupy different mindsets at different times, switching seamlessly between thinking, doing, and making.
- Documenting your thought process, capturing sketches and keeping notes.

- Communication is a two-way relationship: it is about speaking *and* listening.
- Attention to detail is an obligation: the integrity of your work is paramount.
- Do not be precious, have the discipline to not do things, to kill ideas, to avoid scope creep.
- Use reflective learning to improve your capabilities and to make the process work for you.

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

Part B

The Hidden Thinking



3

Formulating Your Brief

In Chapter 2 we learnt about the importance of adopting a process to tackle data visualisation challenges through an organised sequence of activities. Supplemented by the guiding benefit of design principles, this offers a framework to help you make good decisions.

This third chapter initiates the design process commencing with stage 1. This stage encompasses activities concerned with 'formulating your brief' to forge initial clarity about the context and vision of your work.

The manifestation of a brief can be as informal or as formal as your situation requires. For example, it can be useful when working with other stakeholders to document information about the requirements and conditions of a project. It can then be shared, agreed upon and referred back to. It will be in the interests of all parties to have such a source of mutual understanding, especially for matters to do with the expected deliverables. For more personal projects you might need to make basic notes to capture your thoughts.

The primary task of this stage is to establish why you are producing this data visualisation. What is its *raison d'être*? This involves identifying the origin curiosity that will drive your work. This is an articulation of the appetite for understanding you are addressing through your visualisation. No visualisation project is ever undertaken free of constraint, so you will also spend time defining the influential contextual matters around the who, the where and the when. These requirements and factors will shape the conditions of the project you are about to undertake and need to be recognised early.

You will then switch focus, looking ahead to consider the vision of your work. Thinking about this vision represents early conceptual thinking about what it is you might be developing, providing early clues about the best-fit tone, functional experience and style your visualisation may need to demonstrate. The design-centric specifics of how you will fulfil this will be kept on ice until we reach stage 4 later in the process.

In defining your project's purpose, you will consider more deeply what it is for: what are you trying to accomplish? The type of understanding you are facilitating is important. For example, are you imparting key messages to your audience or enabling them to make their own discoveries? Are you placing an emphasis on the precision of readability or amplifying the feeling of a subject? We will also look at harnessing instinctive ideas that form in our minds, concerning the keywords, imagery, metaphors and external inspiration that might be relevant to the subject.

3.1 Defining Your Project's Context

What Is the Motivating Curiosity?

Answering the question 'Where does a data visualisation process start?' might seem straightforward. Instinctively, one might suggest it starts with a request. Someone asks someone else to do a visualisation. They share some background information about the requirements, maybe provide access to some data, and this sets the process in motion.

> 'Be curious. Everyone claims she or he is curious, nobody wants to say "no, I am completely 'uncurious', I don't want to know about the world". What I mean is that, if you want to work in data visualisation, you need to be relentlessly and systematically curious. You should try to get interested in anything and everything that comes your way. Also, you need to understand that curiosity is not just about your interests being triggered. Curiosity also involves pursuing those interests like a hound. Being truly curious involves a lot of hard work, devoting time and effort to learn as much as possible about various topics, and to make connections between them. Curiosity is not something that just comes naturally. It can be taught, and it can be learned.' **Professor Alberto Cairo, Knight Chair in Visual Journalism, University of Miami, and Visualisation Specialist**

This type of scenario is clearly commonplace. However, though this might be *how* a process starts, it is not quite representative of *where* things truly start. You see, before a request is issued, before any data is shared and certainly before any design work is commenced, a *curiosity* has formed: some origin interest held by someone about a subject.

The dictionary definition for curiosity is 'possessing a desire to know or to learn something'. If visualisation is about facilitating understanding, these are the two ends that meet. Curiosity therefore represents the *why* of your process: the instigating, driving motive for a visualisation project to be developed.

You do not create a visualisation because you happen to have data. You create a visualisation because there is a definable appetite for the understanding it offers, whether this appetite is held by you or somebody else (that you essentially inherit). Any visualisation work undertaken in the absence of a definable curiosity will lead to an uncertain and aimless decision-making process.

By identifying your project's origin curiosity, it gives shape to your subsequent decisions, especially those concerned with the content side of your work. This enables you to keep checking that your choices help to contribute towards facilitating understanding about the most relevant matters.

A key attribute of any curiosity is to recognise from whom it originates. Figure 3.1 shows an example of a visualisation produced in response to my own recognised curiosity. This type of work is often characterised as being a 'pet' or 'passion' project that is entirely self-initiated, with no other stakeholder involved. You have freedom to follow your own enquiry, shaped only by the limitations of your imagination and interests. The 'Filmographics' project was entirely motivated by a curiosity I had about the movie industry: 'What are the patterns of success or failure in the movie careers of a range of notable actors/directors?'

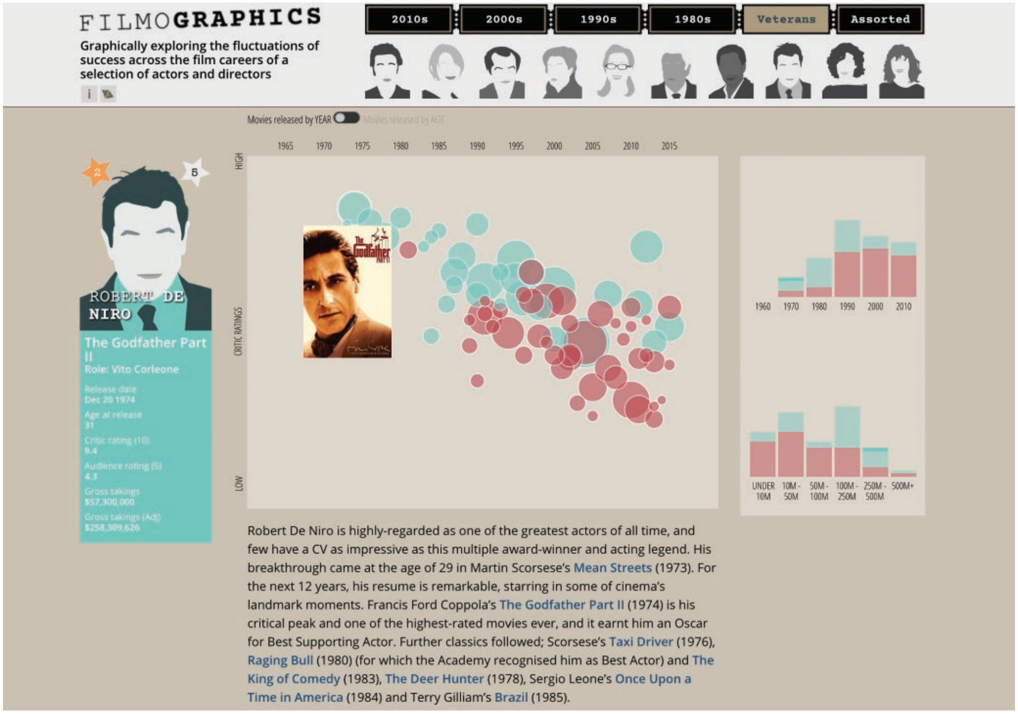


Figure 3.1 Filmographics, by Andy Kirk and Matt Knott

Articulating my curiosity in this way helped focus my decisions about what data to gather, what analysis to conduct and, thereafter, what features of representation and presentation to employ. There were many things I *could* have explored about the movie industry, but this was the particular slice of content that intrigued me the most. I wanted to know about the story of Steven Spielberg's career. I wanted to know whether Meryl Streep's critical successes had been matched by financial success. When was De Niro last consistently making good movies? Why has Adam Sandler been allowed to make *any*?

Although the Filmographics project was initiated by me and served my appetite for understanding, it was published publicly in anticipation of its also being relevant to certain other audiences who perhaps share an interest in the subject matter. I was not explicitly serving their expressed curiosity, rather expecting that some would share my curiosity.

Sometimes the curiosity you are pursuing does not originate from you. Stakeholders are the people involved in a visualisation project, other than yourself, who may influence what curiosity your work should pursue. Stakeholders exist as managers, academic supervisors, clients or colleagues, and it might be them tasking you to undertake a visualisation project based on the curiosity they express to you. In these situations, you are inheriting their interest and you have to own it from then on.

In certain situations, stakeholders may have a dual role of initiator and intended recipient (e.g. 'can you show me trends of sickness absence among staff in my department this year?'), in others they might be expressing to you what they expect a separate audience would find

interesting (e.g. 'can you produce an analysis showing trends of sickness absence among staff this year to share with all heads of departments?').

An important point from the previous chapter needs to be reinforced here: this is the first but not the last occasion when you will have the opportunity, and reason, to refine the definition of your ultimate curiosity. Depending on the particular situation of your work, it may that your initially expressed curiosity is not fixed, not specific and not even singular.

The need for *specificity* in your curiosity will vary from one situation to the next. 'How much did the public engage with the previous Australian election?' is a far broader curiosity than 'What was the percentage turnout across each electoral region of Australia compared with the previous election?' If you are a runner with a fitness-tracking device or application, you might finish a run and wonder 'how good was that run?' This is a broad enquiry. To form an answer requires the synthesis of several distinct pieces of information ('How far? What time? What route? What achievements? What previous times?') that collectively provide a notion of how good the run was. By contrast, if you just want to know 'in what time did I complete the run?', this is a specific curiosity that can be effectively answered by a single piece of information.

On many occasions I have embarked on a visualisation project with an initial curiosity in mind but then, having become better acquainted with the subject through its data, other legitimate enquiries subsequently have emerged as simply being more relevant. It is easier to justify shifting your focus when working on your own projects. When you are being tasked by another stakeholder, you might find there is less room for manoeuvre beyond the pursuit you have been tasked with. That said, you should still seek constant dialogue with your stakeholder if you strongly feel another route might be more interesting. After all, it serves nobody's purpose if you remain anchored to an enquiry that no longer reflects the most relevant aspects of a subject. The important thing to challenge is whether any shift in focus should be embraced or curtailed. No matter how relevant or interesting your new possibilities are, you might simply be drifting beyond the scope of the work.

Sometimes your process is being driven by the need to serve the *known* or *anticipated* interests of your intended audience. If you know your audience well enough and are able to predict their potential needs, your origin curiosity will form around what it is you think *they* want answering. Of course, there will be situations where you are in position to consult your audience directly to define explicitly what appetite they have about the topic in question. Otherwise, you will need to make reasonable judgements to anticipate what *most* people will likely find *most* interesting. These are the situations in which I find most indecision in a visualisation process, mainly because there is always so much choice and so much temptation to mark up everything as being of equal interest. Additionally, at the outset of a project, it might not be reasonable to expect you to be aware of all the potential features of interest about your subject or in your data. This is something that will develop as you work further through stage 2 ('Working with data') and stage 3 ('Editorial thinking').

It is often the case that committing to just a single avenue of curiosity is not feasible. Embarking on multiple distinct curiosities, bound by a shared connection to the same subject matter, might give you more work to do but might be necessary, especially with data or subjects that are unfamiliar to you at the outset.

Suppose you are a student studying the history of music and encounter some data about the structure of popular songs. You suspect there are many potentially fascinating things to discover in this data, but you do not yet know what *the* single most interesting curiosity will be. A key activity of the visualisation process will be to explore the data, unlock its key qualities, test out the multiple different enquiries, and thereafter determine which perspectives offer the most relevant or interesting findings. You are possibly pursuing multiple speculative curiosities in order to see which ones emerge as being most relevant. You might still end up with several, only one, or indeed no legitimate curiosities.

In my experience, forming a question tends to be the most useful and comfortable articulation of a curiosity. In doing this you are positioning your visualisation as a means of providing some notion of an answer. I find this a natural way to keep my mind focused on pursuing this answer.

An alternative way to approach this, especially when you are anticipating what others might find interesting, is to switch your viewpoint away from the question form and think more in terms of what it is you are aiming to present to your audience. You might extend the wording of the data visualisation definition to describe what the facilitated understanding will be about, as illustrated in Figure 3.2.

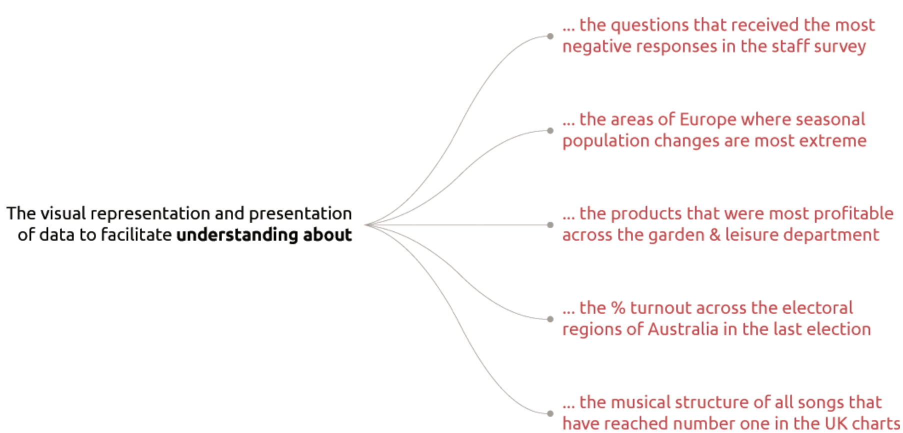


Figure 3.2 Sample Statements of Curiosity

Regardless of your curiosity's origin, specificity, permanence and form, you just need somewhere to start from. Seek to express the most useful overriding curiosity that best encapsulates, at this stage, what you are setting out to pursue for you and/or your audience. Define now, refine later.

Identifying Project Circumstances

The second aspect of contextual thinking concerns identifying a project's circumstances. These are the frictions and freedoms that are imposed *on* you or determined *by* you. Whether you are a full-time visualisation professional, a student, a researcher, working in a business or doing

'Context is key. You'll hear that the most important quality of a visualisation is graphical honesty, or storytelling value, or facilitation of "insights". The truth is, all of these things (and others) are the most important quality, but in different times and places. There is no singular function of visualisation; what's important shifts with the constraints of your audience, goals, tools, expertise, and data and time available.' **Scott Murray, Principal Learning Scientist, O'Reilly Media**

visualisation as a pastime, there are common influencing factors that characterise the conditions of the projects you are undertaking. They will determine the boundaries of your creative ambitions.

When commencing a project, probably not all of the circumstances that may potentially influence your work will be definable. Things change. That is why we need to be prepared to accommodate elegantly the impact of new factors at any point in the process. Of course, the more things you *can* define, the more things become fixed and this reduces uncertainties. Ideally, we want to eliminate as much of the unknown as we can. Useful definition also exists through identifying the absence of restriction or requirement. Knowing you have freedom to determine choices yourself is of clear value. It gives you control. Sometimes you might see merit in imposing restrictions on yourself, where none exist, to aid your focus. Constraints are not always a bad thing; indeed, they can often help us innovate.

People

Stakeholders: In project situations where you have been requested to develop a visualisation by somebody else, it is important to establish an understanding of who is involved. 'Who is the ultimate customer?' is a key question to answer. The customer will always be particularly invested in what you develop. This may or may not be the person(s) who has directly commissioned you, but they will usually determine whether the work is on the right track, in their view. They will also critically determine when the work is ultimately of sufficient quality to be considered finished. In my world as a freelance design consultant, my customer determines when (and if) I will get paid.

Other stakeholders might exist as subject-matter experts, available to offer advice about domain-specific queries. They might be able to guide you on what the most salient issues are in a subject. They might be points of contact to raise issues around technology requirements or advise on some of the nuances you encounter around values held in datasets. On rare occasions, some stakeholders can hinder progress by unduly influencing design decisions beyond their remit and capability. They can become interferers, making conditions harder for you.

Identifying this cast and crew of people who have a stake in your work will help you anticipate the interactions and relationships you might need to manage: what personalities exist, what help you might exploit, and what obstacles might need navigating. Diplomacy will be required.

If there are no stakeholders, and the project is effectively a solo pursuit, there will be much more autonomy, which can be liberating, but with this comes more responsibility on you to direct all matters yourself.

Audience: There are several characteristics of your target or expected audience that will need careful consideration. Your audience will bring irrationalities and inconsistencies. Identifying their varied traits and accommodating their influence into your decision making is a permanent concern throughout the design process and one that requires clear judgement:

- What is your audience's relationship with the *subject* matter? What knowledge do they have or, conversely, lack about a subject? What assistance might they need to interpret the meaning of the subject? Do they have the capacity to comprehend what it means to them?
- What is their *motivation* for acquiring the understanding you intend to provide for them? Do they have a direct, expressed need or are they more passive and indifferent? Might you need to find a way to persuade them or even seduce them to engage?
- What are their visualisation *literacy* capabilities? Might they require assistance perceiving the chart(s) produced? Are they sufficiently comfortable with operating features of interactivity? Do they have any visual accessibility issues, such as red–green colour blindness, that will need to be factored into your design thinking?
- As you will discover, there are usually lots of textual elements included in any visualisation. What regions of the world do your audience come from and, therefore, what *language* considerations must you take into account? Might you need to create multiple translated versions of the eventual solution?

Sometimes, you have direct knowledge of your audience and can easily characterise their needs and mould your choices accordingly. For example, if you have a fixed group of viewers who you know will understand the technical context of the data you are presenting, you probably will not need to include the detailed explanations that would be necessary for a less knowledgeable audience. On other occasions, your audience's characteristics may be more ambiguous and so distant from you that you can only rely on reasonable imagination. You might form in your mind estimated personas of the types of people you could expect to be the main beneficiaries. If you have an especially wide-ranging and diverse profile of audience characteristics, you are unlikely to be able to satisfy the needs of each variation; you might need to commit to prioritising some audiences over others. One size does not fit all.

Visualiser(s): Data visualisation design is truly a multidisciplinary endeavour. It is this variety that fuels the richness of the subject and makes it a particularly compelling challenge. To master it requires a repertoire of skills, knowledge and different attitudes that dominate different stages of this process. Inspired by Edward de Bono's *Six Thinking Hats* (1985), the seven hats of data visualisation in Figure 3.3 represent my attempt to deconstruct the specification of an imagined 'perfect' visualiser. The attributes listed under each of these hats can be viewed as a wish list of personal or team capabilities, depending on the context of your data visualisation work.

> 'There is not one project I have been involved in that I would execute exactly the same way second time around. I could conceivably pick any of them – and probably the thing they could all benefit most from? More inter-disciplinary expertise.' **Alan Smith OBE, Data Visualisation Editor, *Financial Times***

DIRECTOR | The coordinator, overseeing the project

Initiates and leads on gathering and understanding requirements
Identifies and establishes the project's key circumstances
Defines the purpose of the project based on desired outcome
Manages progress through the process and keeps it cohesive
The primary decision maker, often needing to compromise
Pays strong attention to detail
Gets things done: checks, tests, finishes tasks

COMMUNICATOR | The broker between all people

Helps to define the perspective of the audience
A good listener with the humility to defer to domain experts
Has a 'thick skin': needs patience, empathy and diplomacy
A confident communicator with laypeople and non-specialists
Possesses strong copy-editing abilities
Manages expectations and presents possibilities
Launches and promotes the final solution

JOURNALIST | The reporter, pursuing the scent of enquiry

Driven by a desire to help others understand
Defines the origin curiosity of the project
Has an instinct to research, learn and discover
Possesses or is able to acquire salient domain knowledge
Understands the essence of the subject's data
Has empathy for the interests and needs of an audience
Defines the editorial angle, framing and focus

DATA ANALYST | The wrangler, handling the data work

Has strong data and statistical literacy
Possesses technical skills to acquire data from multiple sources
Examines the physical properties of the data
Undertakes initial descriptive analysis
Transforms and prepares the data for its purpose
Undertakes exploratory data analysis
Has database and data modelling experience

SCIENTIST | The thinker, providing scientific rigour

Brings a strong research mindset to the process
Understands the science of visual perception
Understands visualisation, statistical and data ethics
Understands the influence of human factors
Verifies/validates the integrity of all data and design decisions
Demonstrates a systems thinking approach to problem solving
Undertakes reflective evaluation and critique

DESIGNER | The conceiver, providing creative direction

Establishes the initial creative pathway through defining purpose
Harnesses initial mental visualisations: ideas and inspiration
Has strong creative, graphic and illustration skills
Understands the principles of user interface design
Is fluent with the full array of possible design options
Unifies the decision making across the design anatomy
Has a relentless creative drive to keep innovating

TECHNOLOGIST | The developer, constructing the solution

Possesses a repertoire of software and programming capabilities
Has an appetite to acquire new technical solutions
Possesses strong mathematical knowledge
Can automate otherwise manually intensive processes
Has the discipline to avoid feature creep
Works on the prototyping and development of the solution
Undertakes pre and post-launch testing, evaluation and support

Figure 3.3 The Attributes that Comprise the 'Seven Hats of Visualisation Design'

Across these capability groups, which attributes do you possess, or can you demonstrate? Where are your weaknesses, in terms of both gaps and potentially overly dominant traits? I am painfully aware of the things I am simply not good enough at (programming), the things where I rely on instinct more than skills gained from training (graphic design) and also the things I do not enjoy (finishing, proof-reading, note-taking). If certain skills are not available to you compromises may be required and ambitions may need to be lowered.

If collaboration is possible, there are clear advantages in pooling diverse capabilities into a shared challenge. The best functioning visualisation teams will offer a balanced blend of skills across all these hats. Success will be hard to achieve if a team comprises a dominance skewing the diversity of abilities, so what is the best way to allocate or occupy different duties to optimise your design process with a team?

Constraints

Timescales: The primary constraint is usually how much time there is to develop your solution. Most projects have a deadline attached to them, whether this is imposed by other stakeholders, mutually agreed or set by yourself. Even if you do not need necessarily to adhere to a deadline, let's say for personal projects, it can still be useful to define a target date to help sharpen your progress. At the opposite end of the timeline, there is the start date. This may not be *now*. You may have to wait for certain conditions to be in place before you can even commence your work. If you are conducting an analysis of some survey results, you will not have a complete, final dataset of responses to work with until the survey is closed.

During your project there may be certain milestones to factor in as well. These might be occasions when you need to show work in progress or critical points when you switch to working with real data rather than sample data that may be used to draft early ideas.

The most crucial aspect of time is task duration. There are clearly going to be large differences in the ambitions of a project to be completed in two days compared with another in two months. But if the two-day deadline concerns a small-scale task that will take only a few hours, that is going to be deliverable. Though a two-month deadline sounds great, if you are facing three months' work it will be a struggle to accomplish it in time.

Estimating project duration to any reliable degree is a difficult thing to judge. You usually do not know how long a project will take until it is completed, which is often too late to be useful. Even with experience from working on a diverse range of projects, seemingly similar projects can end up with very different task durations. I would always recommend noting down the

> 'What is the *least* this can be? What is the minimum result that will 1) be factually accurate, 2) present the core concepts of this story in a way that a general audience will understand, and 3) be readable on a variety of screen sizes (desktop, mobile, etc.)? And then I judge what else can be done based on the time I have. Certainly, when we're down to the wire it's no time to introduce complex new features that require lots of testing and could potentially break other, working features.' **Alyson Hurt, News Graphics Editor, NPR, on dealing with timescale pressures**

duration of each major task across your design process, so you can more surgically evaluate how you have spent your time and be better placed to estimate accurately commitments on future projects.

Pressures: Depending on the context of your project, certain cost factors may exist. Going back to the matter of time, how much can you afford to spend on a project? Some projects may have a budget allocated and so the associated staff activity costs need to be managed sensibly. You might need to outsource to external parties with specialist expertise, for example transcription services, third-party data sources, illustration work, but can you afford the costs involved? What costs will be incurred in paying for hardware, software, licences to use photography or audio?

Further pressures may emerge from the politics surrounding your subject, the data, or the messages coming from this data. I have been involved in several projects where the charts produced showing data about cities or countries had to be sorted alphabetically, and not by any other ranking measure, in order to preserve a certain diplomatic neutrality. You may receive guidance from your stakeholders that certain messages need to be downplayed or amplified. These can be difficult matters to handle: you want to respect any requirements received, but also you do not want to undermine the integrity of what you are representing.

There may be cultural sensitivities to consider if you are creating work for audiences from different regions. Issues around the use of imagery, colour connotations, or symbology of certain forms may need to be carefully handled. There may also be environmental considerations, particularly concerning the output of your work, that need to be observed.

Design: Restrictions around certain design choices are common, often informed by style guidelines that must be adhered to through the use of specific colours, typeface and fonts. Where possible, I always attempt to challenge these rules somewhat because they can be unnecessarily restrictive, but permission is not always granted. You may need to include logos, which can take up valuable space and unbalance your composition, but rules are rules and therefore we need to know about these things at the outset, not later.

Layout or size restrictions may also exist, dictating the space in which you have to work. For example, when producing graphics for journals or for digital outputs that need to work on a tablet or smartphone, you might have quite a small amount of space to utilise. Conversely, your output might need to be very large, which can introduce different challenges with legibility and resolution quality.

Further creative pressure might materialise in the form of what I describe as market influences. The visualisation you develop may have to compete for attention alongside other work. For example, if you are creating a visualisation for a charitable organisation, how do you get a message across louder and more prominently than others competing for the same eyeballs? If you are working on an academic research project, how do you get your findings heard among all the other studies battling to create an impact? Creative influences can emerge internally, through the unique dynamics of an organisation, and externally, through broader competition across the entire marketplace and industries. Although it is not the most important factor, a desire to emulate the best or differentiate from the rest can prove to be a strong motive in your design thinking.

Technological: As I have mentioned in the Introduction, there are myriad tools, applications and programming libraries in data visualisation, offering a varied landscape of capabilities. The technology you have access to will affect how digitally ambitious your work can be and/or how efficiently you will be able to make it. You can only achieve what your tools enable you to achieve. This influence will shape several stages of your process:

- *Working with data:* Technologies to help with acquiring, examining, transforming and exploring data. How much data can your tools handle? How quickly do they perform actions, especially with large data? Do they enable automation, maybe through scripting? What range of statistical techniques is available? How effective are the methods for modifying data? Do they enable you to explore your data visually?
- *Data representation:* Technologies to help with making charts. What range of different charts do they enable you to make? Is the process of constructing charts automatic or manual? Do the tools facilitate workarounds or means for potentially expanding their standard capabilities?
- *Interactivity:* Technologies to develop features for exploration and control. What range of different interactive features do they offer? Is the process of developing these features automatic or manual? Do the tools facilitate workarounds or means for expanding their potential capabilities?
- *Data presentation:* Technologies to manage the inclusion of annotations, the use of colour and the composition of your work. What range of annotated features can you include? Are you able to control fully the appearance of these features? To what extent can you manage the colour applied to every visual element? Likewise, what degree of control do you have over the size and placement of all elements?
- *Publishing:* Technologies to disseminate your work. Do you present through a slide deck? Do you compile a printed report? Do you publish your work through a website? Do you upload it to be accessed/downloaded from the Web? Do you send files via email to others with the same applications? Do you compile a video? Do you publish as a gif on social media?

Deliverables

Setting: This concerns the characteristics of the setting in which your work is going to be encountered and consumed by your audience. Firstly, is it going to be consumed remotely – away from you – or presented in person? If you are not personally present to offer verbal explanations of key features and findings, descriptions of the data gathering process, assumptions or calculations, you may need to include these as annotated properties. Secondly, is the nature of the engagement one that needs to facilitate especially rapid understanding, or does it lend itself to a more extended, prolonged engagement? I usually think of four broadly typical settings:

> 'I like to imagine that I have a person sitting in front of me, and I need to explain something interesting or important about this data to them, and I've only got about 10 seconds to do it. What can I say, or show them, that will keep them from standing up and walking away?' **Bill Rapp, Data Visualisation Designer, discussing an audience scenario setting he conceives in his mind's eye**

- The *boardroom*: A setting characterised by limited time, limited patience and limited attention. Immediate insights and key messages need to be imparted at a glance. There will likely be reduced appetite for engaging with anything that requires effort, such as an unfamiliar chart type or rich interactivity, unless someone is there to present it.
- The *coffee shop*: A more relaxed setting that might be compatible with a piece of work that involves more effort and requires more time to learn about the subject. Unfamiliar representations might be reasonable to use as long as sufficient assistance is provided about how to read the displays. Interactive features to enable interrogations may not pose the same obstacles to understanding in the way they would in other situations.
- The *cockpit*: A situation characterised by operational need, whereby the visualisation is offered as a tool or instrument to provide immediate signals that stand out at a glance. Sufficient breadth and depth of additional content will also be required to serve the multitude of different potential scenarios that might arise. Think of a wide-ranging organisational dashboard or a reference map that serves multiple potential levels of enquiry that aid the operational needs of navigation, from high-level orientation to in-depth localised detail.
- The *prop*: Here a visualisation plays the role of a supporting visual device to accompany a presenter's verbal facilitation of the key understandings (perhaps in the form of a talk or video) or an author's write-up of key findings that refer to an accompanying chart or figure (report or article).

Medium: You will need a clear understanding about the specific format of the deliverables required. Is your intended output to be produced for print, for digital or maybe even physical? Will it be static, interactive or animated? An important thing to reinforce again: just because something is published on the Web does not mean it is interactive.

> 'I love, love, love print. I feel there is something so special about having the texture and weight of paper be the canvas of the visualisation. It's a privilege to be able to design for print these days, so take advantage of the strengths that paper offers – mainly, resolution and texture. Print has a lot more real estate than screen, allowing for very dense, information packed visualisations. I love to take this opportunity to build in multiple story strands, and let the reader explore on their own. The texture of paper can also play a role in enhancing the visualisation; consider how a design and colour choices might be different on a glossy magazine page versus the rougher surface of a newspaper.' **Jane Pong, Data Visualisation journalist at the *Financial Times***

Maybe you will have to produce something that will published across multiple media. For example, newspapers typically publish graphics in their printed version, on the web version, on their mobile apps, and often also share them via social media. While it may be the same graphic produced four times over, there may still be subtle alterations required to optimise the presentation for each respective platform, especially when considering the impact of the size restrictions that will exist. This increases your workload.

Having spoken about technology, maybe you do not need any? In the right context, it may be possible to embrace more analogue or artisanal approaches, as demonstrated in the example shown in Figure 3.4, which is a visualisation (or possibly better termed a 'data physicalisation') created using Play-Doh.

This analysis presents data about the various plug types used around the world. The imprints of the configuration of pins for each plug type are stamped into coloured chunks of Play-Doh, and then white lengths are measured out to represent the populations of people from the countries whose power systems use each type.

Quantity: As well as the medium, it is also important to get a sense of the project's deliverables in terms of expected quantities. How many *things* are you making? How much, what type, what shape and what size? For example, are you producing 12 different graphics for a month-by-month slide deck or contributing to a large report that will need two charts developed for each of the 50 questions asked in a survey? Perhaps it's a web-based project with four distinct sections, each requiring at least one interactively adjustable view, or it could be just a single chart to be emailed to your manager. It is not always possible to determine output quantities this early in the process, but you should certainly maintain awareness of how realistic the expected deliverables are going to be, given the project's resources (timescales, skills and budget, where relevant).

Frequency: The issue of frequency concerns how often a particular project will need to be reproduced and what its lifespan will be. It might be a regular (e.g. monthly report) product, in which case the efficiency and reproducibility of your design solution will be paramount. If it is a one-off or irregular piece (e.g. election polling graphic updated after each new release), you will have more justification to create a bespoke solution so long as the cost–benefits involved remain positive. Maybe it is a one-off project in development terms but will be constantly updated and republished on a frequent basis thereafter, such as a monthly report. There may be more upfront work to develop a functionally robust template, though the task of generating each subsequent monthly report may only involve a limited amount of work. You may need to consider if there will be any future benefits from reusing some of the techniques you have employed, so there is some recyclable value. Can you justify investing time, for example, in programmatically automating certain parts of the construction process if they can be reused to save time in the future?

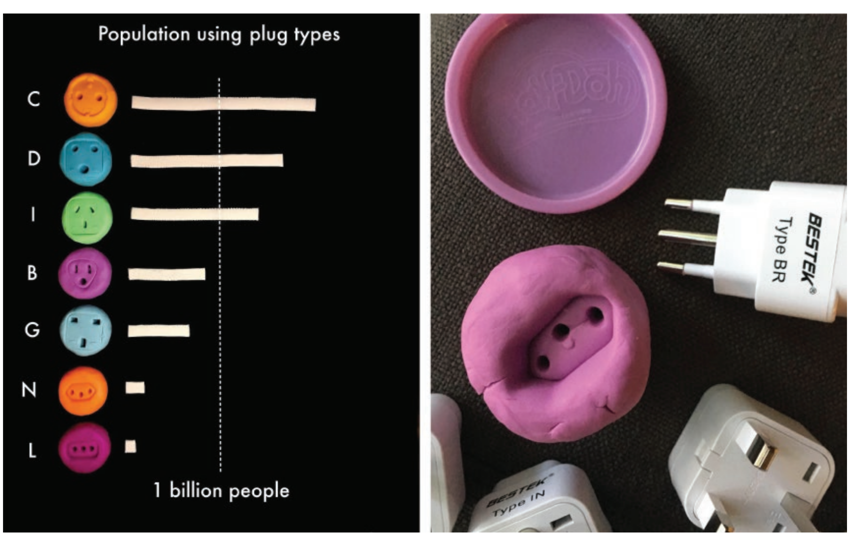


Figure 3.4 Popularity of International Outlets, by Amy Cesal

3.2 Establishing Your Project's Vision

Defining Your Project's Purpose

Identifying the curiosity that motivates your work establishes the project's origin. The circumstances you have just considered define the conditions you will experience and need to accommodate through your project. To supplement this contextual thinking, the vision you have for your work needs some early consideration.

The definition of *vision* is 'the ability to think about or plan the future with imagination or wisdom'. What are you hoping to accomplish with your visualisation: what is its *purpose*?

Articulating your project's purpose represents a statement of intent. It offers clarity about what you see as your destination. As before with the origin curiosity, purpose might evolve as you progress through the process, but the sooner you can establish at least some degree of focus, the better, especially in being able to eliminate potential creative avenues that will have no relevance to your aims.

We have established that the overriding goal of a visualisation is to facilitate understanding, though the *nature* of understanding can vary from one project to the next. Some visualisations aim to be quite impactive, attempting to shock an audience into changing behaviour or perhaps inspiring viewers to take significant action. For example, you may seek to change attitudes among parents about the effect of sugary drinks in contributing to the rise of obesity. To accomplish this might require an emotive style that amplifies the feelings about the subject, attracting the attention of impassive viewers and then striking home with a powerful message that, hopefully, resonates deeply with anyone in a position to act. Affecting people to this extent can be ambitious, but it might be the purpose of the work to be this ambitious.

In another context, a visualisation about obesity may hold more modest ambitions of just seeking to inform viewers about a subject. Let's suppose you are offering health professionals an interface that lets them explore obesity trends in their local area. They probably will not need convincing about the significance of this topic or any aesthetic seduction to encourage them to participate. As health professionals they are likely to have an operational need to know this and a certain responsibility to educate citizens themselves. The most suitable style of visualisation in this case might therefore be more low key, perhaps imparting an authoritative tone with an emphasis on technical precision and clear functionality. If all it achieves is to reinforce existing understanding or just add an extra small grain of understanding, this will possibly represent success. Purpose achieved.

There are different types of visualisations that demonstrate different design characteristics, the suitability of which will be largely determined by what you are trying to accomplish. In Chapter 1, I described how viewers go through three phases of understanding: *perceiving*, *interpreting* and *comprehending*. I explained how, as visualisers, we have limited control over the final phase, comprehending, which is largely determined by a viewer's attitude and connection to the subject matter. We do, though, have control over how our viewers perceive and interpret our visualisations. They are particularly influenced by the choices surrounding two significant design characteristics: *tone* and *experience*.

Judging the Tone of Your Visualisation

The tone conveyed by a visualisation has an influence on the perceiving phase of understanding. In judging the most suitable tone for your project, you are deciding whether to place more emphasis on the viewer being able to read data or feel data (Figure 3.5).



Figure 3.5 The Spectrum of 'Tone'

Reading tone: A visualisation that conveys a *reading* tone places emphasis on optimising the precision and efficiency of perceiving the represented data. The visual quality that embodies a reading tone of voice might be described by adjectives like pragmatic, authoritative, analytical, conservative, utilitarian and (necessarily) boring.

A reading tone might be suitable in circumstances where there is no need to employ any form of visual stimulation to impart a message more potently, nor to seduce an audience through aesthetic appeal. There will be good reason to portray the underlying subject in a statistical style and there is no desire or relevance in amplifying any emotional devices.

The design choices employed will seek to make it easier for a viewer to determine the magnitude of and the relationships between values. Representation methods like bar charts, as shown in Figure 3.6, embody this tone of voice. By representing the size of a quantitative value using the proportional size of a line, bar charts facilitate both general sense-making and precise point-reading, thus heightening the perceptual accuracy for the viewer. In this example you can quickly ascertain that the USA value is about three times the size of the next largest, the UK one. Although similar in magnitude, you can see that the value for Italy is slightly larger than the one for The Netherlands which is slightly larger than those for Canada and Switzerland.

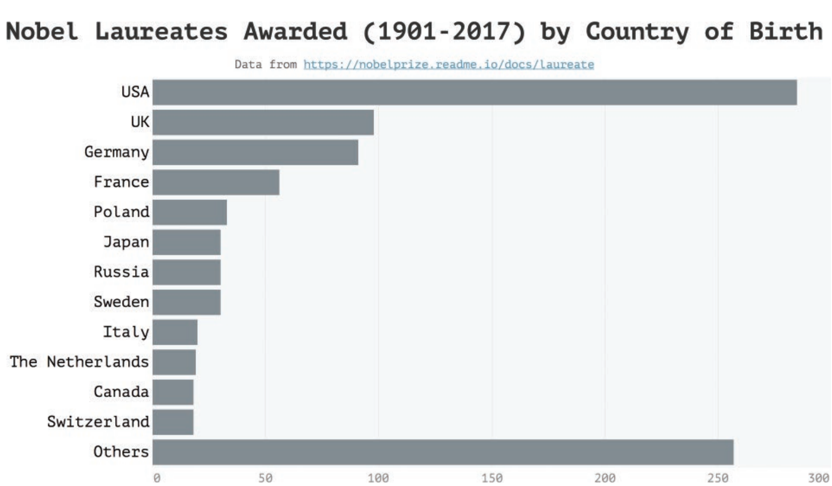


Figure 3.6 Nobel Laureates Awarded (1901–2017), by Country of Birth

You can probably estimate France's absolute value to be around 55 and Germany's around 90. For exactness in reading the values you would offer direct value labelling, though the degree of accuracy in judgements here is already quite high and probably sufficient.

Bar charts are so ubiquitous and so necessary because they make it easier to answer the quite reasonable question: 'What is the size of that value?' Most of the visualisations you will ever produce will likely lean towards offering this kind of *reading* tone.

Indeed, you might reasonably ask why would you ever *not* seek to optimise the accuracy and efficiency of value judgements? Surely anything that compromises on this is undermining the accessibility of your design and maybe even jeopardising its trustworthiness? Well, this is why the definition of your purpose is so significant. There are other considerations, as typified by the *feeling* end of this continuum.

Feeling tone: In contrast to reading values, sometimes we might justifiably place more emphasis on *feeling* data. The visual quality that embodies a feeling tone of voice might be described using adjectives like emotive, seductive, figurative, big-picture, fun and dramatic.

Sometimes, the perceptual judgements that are most important for your viewer may align more with the notion of 'getting the gist'. This means the viewer can quickly and easily form headline observations of the hierarchy of large, medium and small properties of your data. The viewer might gain a general *sense* of major patterns that reveal things going up and going down, the major clusters of connections and the major components of a whole. A representation of data that facilitates 'at-a-glance' viewing is sometimes the most suitable way to portray a subject's values. The consequence of this is that perceiving precise readings is diminished.

As described earlier, the benefit of visually representing data is that it offers something different, and often something better than a table of data, by helping a viewer see quantitative and qualitative relationships in a subject. There are occasions when we want to do more than just let a viewer see the subject through its data. Sometimes you will be working with subject matter that has the potential to stir strong emotions or relates to inherently imprecise or abstract concepts.

The projects displayed in Figures 3.7 and 3.8 exist at the very intersection of these notions, portraying the imprecision of emotion as conveyed through the use of language. The work in Figure 3.7 is an excerpt from a project looking at the emotional arcs of the past ten US presidential inaugural addresses using the Microsoft Emotion API to analyse facial expressions and match them to common emotion classifications. Each 'feather' form represents a full inaugural address and each barb of the feather is a moment during the speech where the president displayed an emotion: positive emotions are drawn above the quill, negative emotions below. The length represents the intensity of the emotion.

Figure 3.8 analyses the emotions found in Taylor Swift's song lyrics. In this project the data was processed using IBM Watson to derive emotions of happy, sad, mad and scared. The emotional mix of each track is then represented like a mix of gooey liquid with different measures of yellow, blue, red and purple. The size of each blob represents a measure of the confidence in how the process has identified explicit emotions; the smaller songs indicate Watson could not absolutely detect outright emotion.

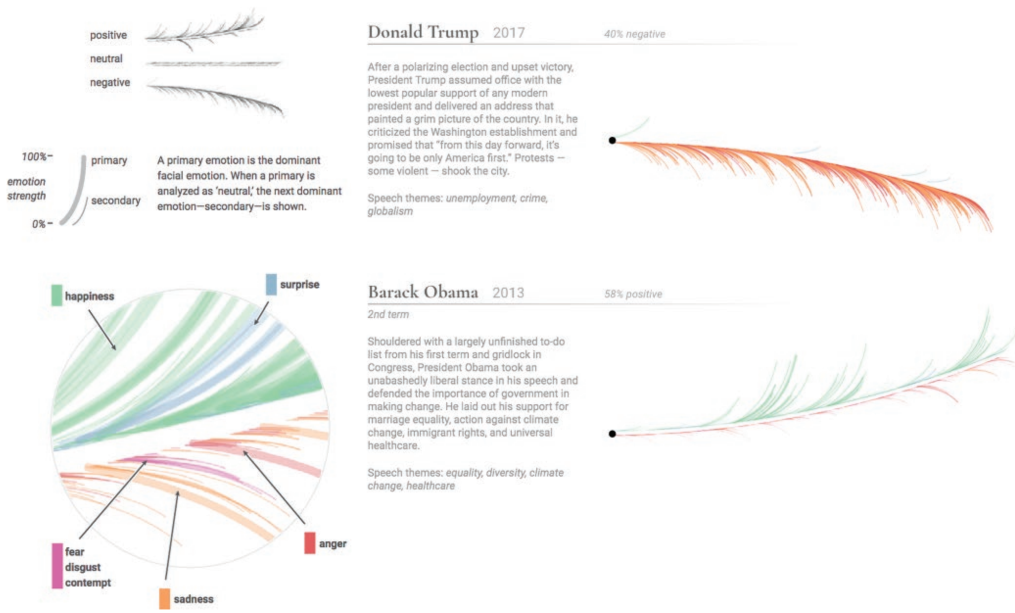


Figure 3.7 One Angry Bird, by Periscopic

The data in both these works is based on a good degree of automated subjectivity, but the respective portrayals perfectly embody the feeling of the subject: you cannot read them with precision and you should not seek to read them with precision, because the data represented in them does not convey precision.

As I mentioned, sometimes we do not need to read values precisely because it is more important just to get the gist. In the project illustrated in Figure 3.9, you see visuals from an analysis of the families who have most financial clout when it comes to funding presidential candidates. The data quantities are portrayed using stacks of Monopoly house pieces piled up outside on the White House lawn. The red houses represent the small number of families who have contributed nearly half of the initial campaign funding, the green pieces are representative of the total households in the USA. You cannot count the number of pieces piled up. You cannot get remotely close to estimating their quantity, but you can get a sense of their relative proportion from the juxtaposition of *many* compared to *few*. It offers a visual approximation of the remarkably disproportionate balance and power of wealth. That is the only level of readability offered and intended.

Data is more than just a bunch of numbers and text values. Thinking about tone is to recognise semantically what your subject is about: what activity, instance or phenomenon does it represent? Is it about people, places, products? Is it about similarities or differences, change or growth?

Learning about the underlying phenomena of your data helps you feel its spirit more clearly than just looking at values in isolation. This prepares you for the level of responsibility and potential sensitivity you will face in curating a visual representation of *this* subject matter.

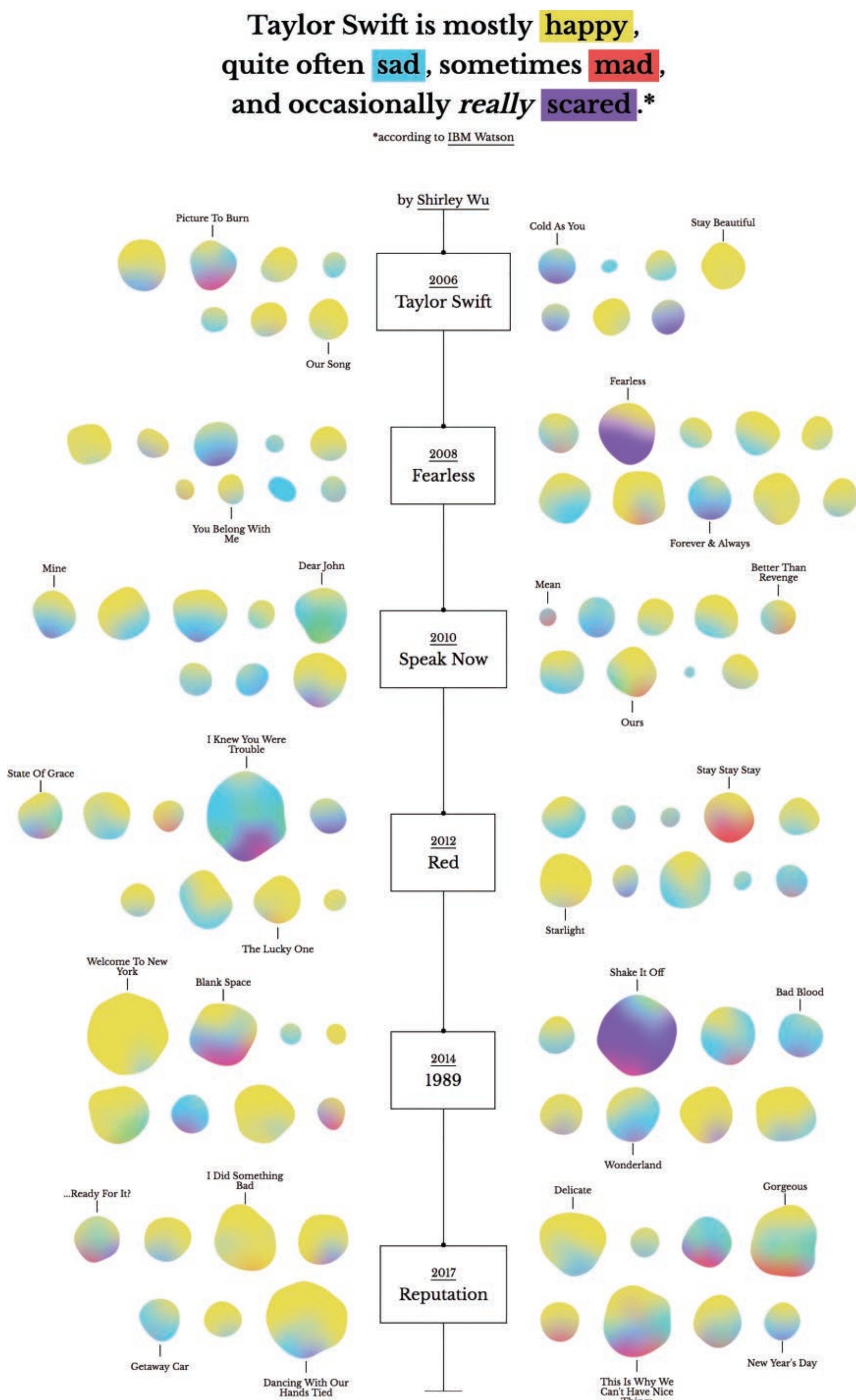


Figure 3.8 Taylor Swift is Mostly Happy, Quite Often Sad, Sometimes Mad, and Occasionally Really Scared, by Shirley Wu



Figure 3.9 Buying Power: The Families Funding the 2016 Presidential Election, by Wilson Andrews, Amanda Cox, Alicia DeSantis, Evan Grothjan, Yuliya Parshina-Kottas, Graham Roberts, Derek Watkins and Karen Yourish (*New York Times*)

As we saw in Chapter 2, with the case of the 'Florida Gun Deaths' graphic (Figure 2.5), some subjects are inherently more emotive than others. You might choose to amplify or suppress the emotion of the subject, and you need a clear conviction in deciding how to find the most suitable tone of voice.

> 'There's a strand of the data viz world that argues everything could be a bar chart. That's possibly true but also possibly a world without joy.' **Amanda Cox, Editor, *The Upshot***

For subjects that carry the weight of strong emotion, there might be good reason to exploit the inherent feelings. Encapsulating emotional sensations like fear, disgust, fun and humanity through your design choices might accelerate the meaning of the subject and potentially affect the most elusive phase of understanding, comprehending and how viewers feel.

This approach could be seen as somewhat manipulative. To a certain degree it probably is and there are risks associated with misjudging the employment of emotional attributes. A playful approach to portraying data about a serious topic will demonstrate insensitivity and possibly undermine the trustworthiness of your work, even if you have created an elegant solution. As long as you are faithful to the underlying data and the subject's visual embodiment is not superficial, artificial or deceptive, I believe it is an entirely appropriate motive when the circumstances suit.

> 'Find loveliness in the unlovely. That is my guiding principle. Often, topics are disturbing or difficult; inherently ugly. But if they are illustrated elegantly there is a special sort of beauty in the truthful communication of something. Secondly, Kirk Goldsberry stresses that data visualization should ultimately be true to a phenomenon, rather than a technique or the format of data. This has had a huge impact on how I think about the creative process and its results.' **John Nelson, Cartographer**

It is important to note that any visualisation work that leans more towards ‘feeling’ is typically the exception and will be relevant to a minority of situations. A skilled visualiser needs an adaptive view and the ability to judge the appropriate occasions whereby the purpose of visualisation will support such an exceptional approach.

When it comes to defining the best-fit choice of tone, it is often possible to think of a blend of options in combination. There will be projects you work on that involve multiple chart assets, multiple interactions, different pages and deeper layers. The mantra proposed by Ben Shneiderman (1996), one of the most esteemed academics in this field, namely ‘Overview first, details on demand’, informs the idea of thinking about different layers of readability and depth in your visualisation work accessed through interactivity. Some of the chart types that you will meet in Chapter 6 can only ever hope to deliver a *gist* of the general magnitude of values (the big, the small and the medium) and not their precise details. A treemap, for example, is never going to facilitate the detailed perceiving of values. In the example shown in Figure 3.10, showing S&P 500 stock, the area of each rectangular shape represents the size of market capital for each company included. The colours indicate the change over the past 24 hours.



Figure 3.10 Finviz: Standard & Poor’s 500 Index Stocks (www.finviz.com)

Our perceptual system is quite poor at estimating scales of areas, as you will learn later. If you wish to compare the size of one stock (e.g. Google, top left) with another (e.g. Amazon, top middle) it will not be easy to make accurately such a judgement. However, you do get a sense that they are both relatively large and a chart like this usually seeks only to give a general sense

of the hierarchy of values (big, medium, small) as well as prominent observations of colour (vivid red vs vivid green). The clue is perhaps in the name – tree*map* – in that some charts often provide multiple layers of detail, navigating from a broad understanding of how complex or dense a system of content towards more detailed specific enquiries thereafter. In this case, as you can see, features of interactivity exist allowing the user to hover over a given shape to reveal a tooltip containing precise details as value labels.

In the Better Life Index, shown in Figure 3.11, the initial view is based around a series of charts that look like flowers. This is attractive, intriguing and offers a nice single-page summary at a glance. The task of reading the petal sizes with any degree of precision is hard but that is not the intent of this first layer. The purpose is to achieve a balance between a form that attracts the user and a function that offers a general sense of where the big, medium and small values sit within the data. For those who want to read the values with greater accuracy, once again, they just need to hover over the flower shapes to view an alternative representation of the same data in the form of a bar chart.

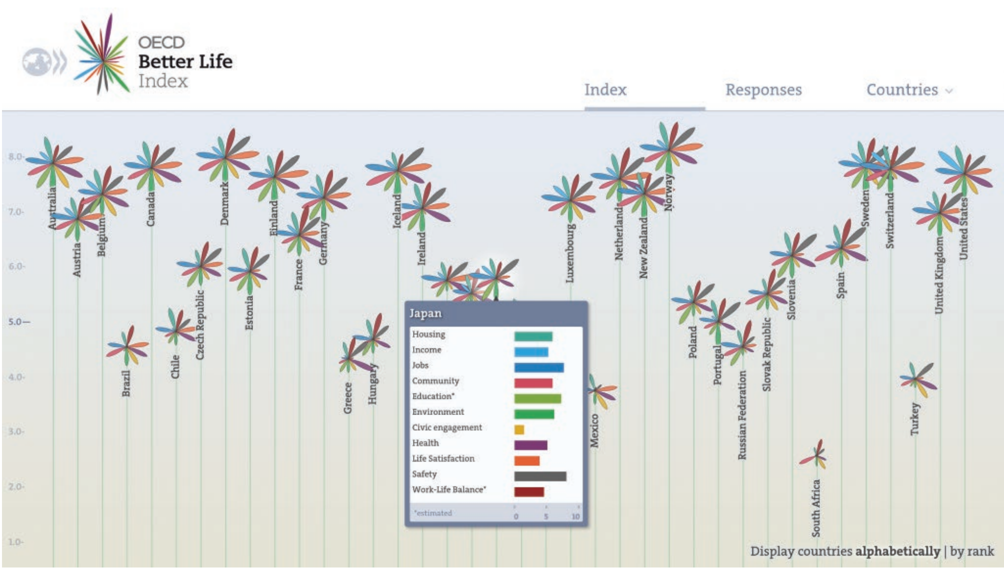


Figure 3.11 OECD Better Life Index, by Moritz Stefaner and Dominikus Baur, Raureif GmbH

In both these examples the viewer's task of perceiving the chart has been adapted from a general feeling of the data towards a more precise reading of the values. Sometimes this initial 'gateway' layer is required as a primary view, to seduce your audience and/or to provide a big picture at a glance (feeling), and then the audience move towards more perceptually precise displays of data (reading). This is usually achieved through interaction or by any means of sequencing, perhaps by navigating through the pages of a report or advancing through a slide deck.

Judging the Experience Offered by Your Visualisation

The experience offered by a visualisation influences the interpreting phase of understanding. Whereas tone embodies a continuum, the judgement of the most suitable experience is more distinct and concerns different methods of enabling interpretation: *explanatory*, *exhibitory* or *exploratory* (Figure 3.12).



Figure 3.12 The Classifications of 'Experience'

Explanatory visualisations offer an experience characterised by the visualiser taking responsibility to present important observations and interpretations to help the viewer more quickly assimilate the meaning of what is presented. I find *quotation marks* are emblematic of explanatory visualisations, as they are associated here with a visualiser saying something.

The simplest, perhaps mildest method of creating an explanatory experience is through the inclusion of simple devices that direct the eye's attention towards key features of a display. Visually emphasising values of most interest through the use of contrasting colour properties can offer cues that establish the hierarchy of importance. Annotation properties like value labels, captions or summaries can provide explicit textual commentary about key findings to accelerate the interpretation.

An example of this kind of explanatory experience is shown in Figure 3.13, which was published in an article reporting on protests across US schools (in November 2015) regarding the under-representation of black students.

Here you see a scatter plot comparing the share of enrolled black students for different public research universities (along the vertical y-axis) with the share of the college-age black populations in the respective states (along the horizontal x-axis). With the protests beginning at the University of Missouri, the chart uses red to highlight this data point as the primary item of interest. Other notable colleges, as mentioned in the article, are emphasised using darker dots and labels to illustrate useful comparisons. With additional visual overlays like the trend line and dotted line indicating proportional representation, the viewer's attention is drawn to the implication of what it means to be positioned in different regions of this chart – is it good or bad, typical or atypical?

A useful way to consider the role of an explanatory visualisation is to think how you would verbally present key insights from any chart in person. What features would you point out as being the most interesting? Which values would you mention, and which would you ignore? The traits of a good explanatory visualisation are that it effectively does the job of communicating the main features you would remark on if you were there. It can stand alone without the need for in-person explanation, yet still beckons the viewer towards important interpretations.

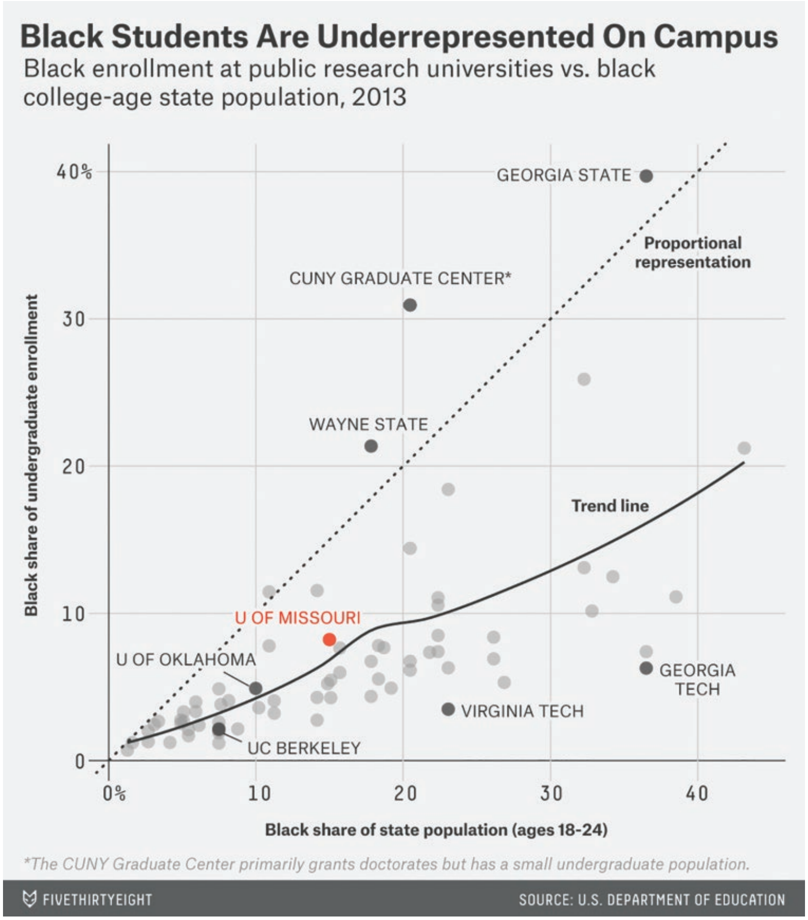


Figure 3.13 Mizzou's Racial Gap is Typical on College Campuses, by FiveThirtyEight

A more intensive example of an explanatory visualisation would be characterised by work that enlightens through narrative sequences in the form of sophisticated articles, animations or presentations. Some might describe this as 'narrative' visualisation. This is where the most tangible demonstration of storytelling is relevant. One example that typifies this classification is seen in a powerful video illustrated in Figure 3.14 through a selection of still images. The video employs an animated graphic sequence to weave together a data-driven narrative describing issues of wealth inequality in the USA. There is an affecting voiceover that verbally presents the main insights of the subject at hand, delivered via a linear story that unfolds. As a viewer, you sit back, listen and process what you are being told.

Common to any explanatory visualisations is a need for the visualiser to possess sufficient knowledge – or have the skill and capacity to acquire sufficient knowledge – about the topic being displayed. The visualiser needs to be able to identify the most relevant and interesting insights to present to the viewer. Creating explanatory visualisations forces a visualiser to challenge how well he or she actually knows a subject. If you cannot explain or articulate what is insightful, and why, to others, then this probably means you do not know the reasons yourself.



Figure 3.14 Excerpts from Wealth Inequality in America, by YouTube user 'Politizane'

Explanatory projects will mainly be for audiences who do not have the knowledge, capacity or time to form for themselves the meaning of a visualisation. Furthermore, if you have *something* to say, indeed if you *have* to say something, say it with an explanatory visualisation.

Exploratory visualisations differ from explanatory in that they are focused more on helping the viewers or – more specifically in this case – the *users* discover and form their own interpretations. Almost universally, these types of works will be digital and interactive in nature. I find the *question mark* is emblematic of exploratory visualisations, as they are associated with a visualiser helping a user answer a question.

The most basic level of exploratory visualisation provides simple interrogation and manipulation of the data. You might offer your user the ability to filter a display to show only certain categories of interest or switch the view to different data parameters.

An example of this type of visualisation is shown in an interactive project in Figure 3.15. It was developed to allow users to explore different measures concerning the dimensional changes of different wood species, over time, across selected cities of the world. There are no captions or conclusions. There are no indications of what is significant or insignificant. There is no assistance from the visualiser to help the user interpret the meaning of this data – what is 'good' or 'bad'? This project exists simply to provide a visual window into the subject through this data to enable users to interact with the different indicators and selections offered to let them find features that resonate and form their own interpretations.

The responsibility for then translating 'what it means', the essence of interpretation, is passed to them. This kind of experience will only be suitable if the audience have the requisite knowledge and motivation to form such interpretations themselves. Indeed, the assumption would be that the users will be better equipped to do this than the creators.

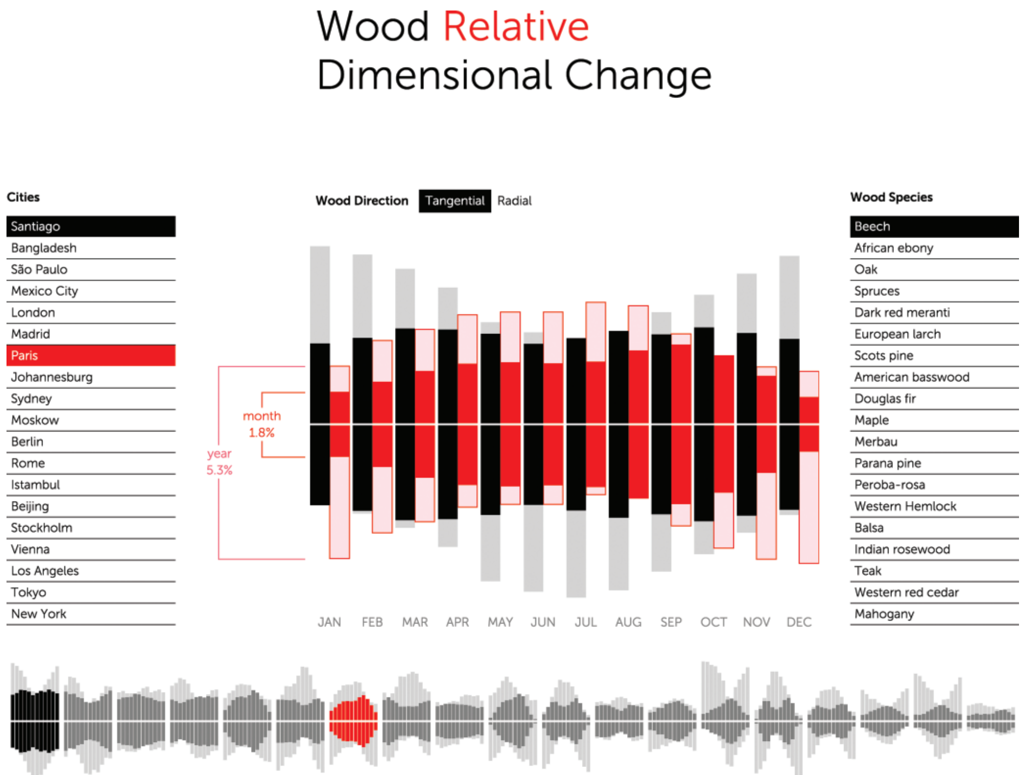


Figure 3.15 Dimensional Changes in Wood, by Luis Carli (luiscarli.com)

A deeper exploratory experience goes beyond just offering means to interact and more towards what might be described as offering a participatory or contributory experience. The prospect of greater control, a deeper array of features and the possibility of contributing one's own data to a visualisation can be very seductive. Users are naturally drawn to challenges like quizzes and projects that allow them to make sense of their place in the world (e.g. how does my salary compare with others; how well do I know the area where I live?). Figure 3.16 shows the *New York Times*' so-called 'Dialect quiz map'. This is just one contemporary example employing this participatory approach to great effect.

In this case users are invited to complete 25 questions about their use of language terms in different scenarios. Based upon their responses and the others gathered in the associated (and ever-growing) study, the similarity or otherwise of their apparent dialect compared across the USA is revealed graphically. This is a custom map display shaped by the contributions of the participating users. It shows *them* who *they* are. You might think this outcome is more characteristic of an explanatory experience, but the end state is only reached as a result of the user's participation. And even then, it is down to the user to interpret the meaning of the results shown. There was not any one thing the visualiser wanted to say, rather there were many thousands of things – the only way to impart this was by handing over control to the users to let them discover for themselves.

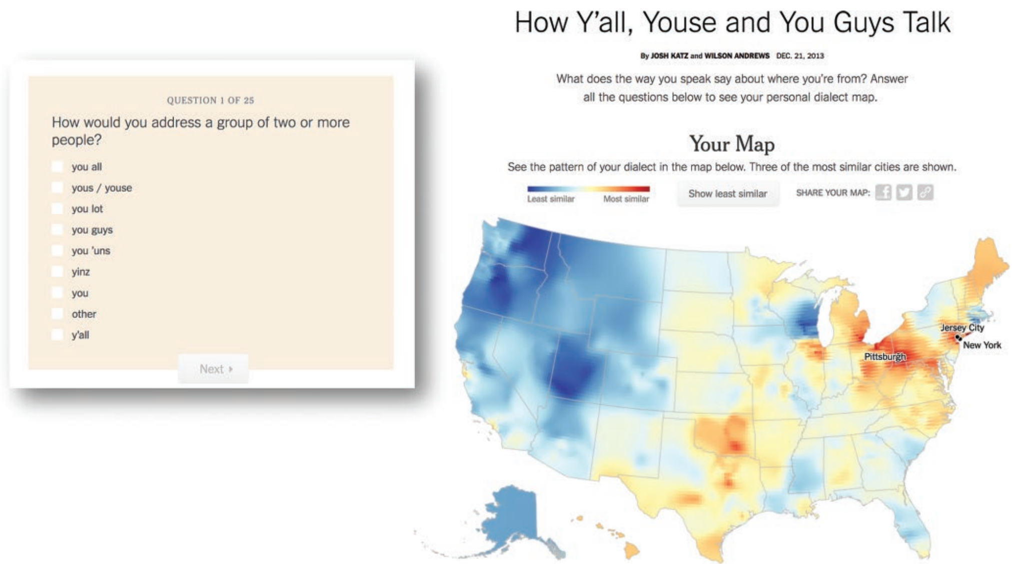


Figure 3.16 How Y'all, Youse and You Guys Talk, by Josh Katz (*New York Times*)

The biggest obstacle to the success of an exploratory visualisation's impact is the 'so what?' factor. 'What do you want me to do with this project? Why is it relevant? What am I supposed to get out of this?' If these are the reactions you are getting from users, then there is a clear disconnect between the intentions of your project and the experience (or maybe expectations) of those using it.

Increasingly there is a trend for visualisation projects to blend different types of experiences into the same overall project – the term 'explorable explanations' has been coined to describe them. A project like 'Losing Ground' by ProPublica (Figure 3.17) is an example of this as it moves between telling a story about the disappearing coastline of Louisiana and enabling users to interrogate and adjust their view of the data at various milestone stages in the sequence.

Exhibitory visualisations are characterised by being neither explicitly explanatory nor functionally exploratory. With exhibitory visualisations the viewers have to do the work to interpret meaning, relying on their own capacity to perceive and translate the features of a visualisation. I generally describe these visualisations as simply being visual displays of data and find the *ellipsis* is emblematic of exhibitory visualisations, as it represents the idea of a visualiser leaving the viewer to finish the task of gaining understanding.

Think of this type of experience in relation to exhibiting an artwork: it takes the interpretative capacity of the viewer to be able to understand the *content* of a display as well as the *context* of a display. In contrast to exploratory visualisation, for exhibitory pieces this is conducted just by looking and thinking. But like exploratory experiences, exhibitory projects rely entirely on the audience having the motivation and capacity to interpret.

You might wonder what the value is of an exhibitory visualisation. Sometimes the circumstances of the audience encountering a visualisation do not require technical exploration or direct

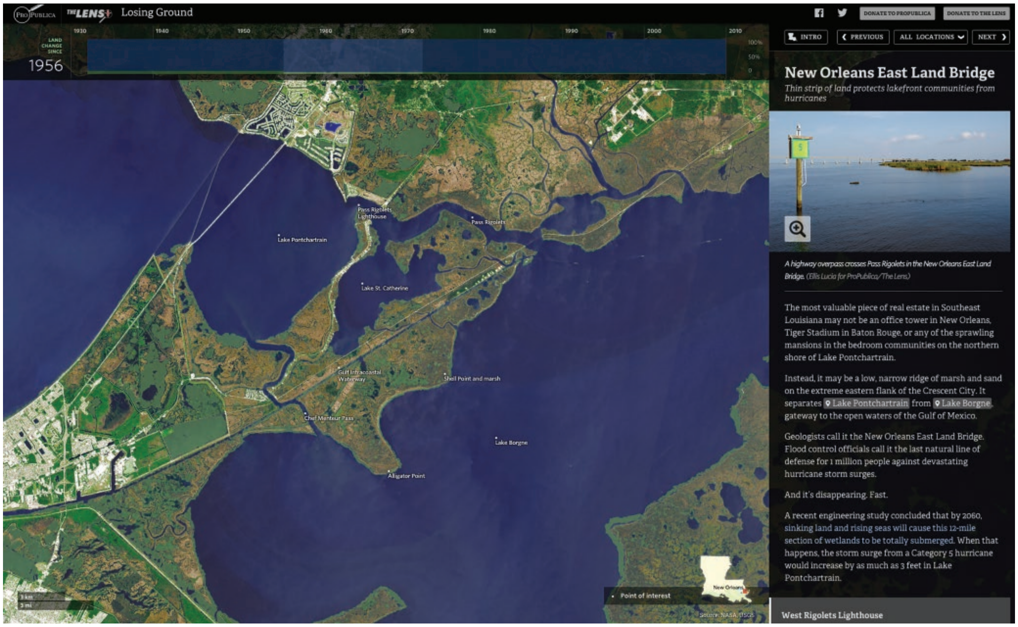


Figure 3.17 Losing Ground, by Bob Marshall, The Lens, Brian Jacobs and Al Shaw (ProPublica)

explanation. If you have a very specific audience whom you know to be sufficiently knowledgeable about the domain and the analysis you have provided, it might not be necessary to emphasise any of the key insights as you would with an explanatory visualisation. Furthermore, the extent of the analysis might be so narrow that there is no value in enhancing the experience with interactivity. Indeed, it might not even be technically feasible to do so.

In Figure 3.18, there is an analysis of the top 100 highest earning athletes. This is an exhibitory piece because it leaves you to form your own interpretation about what you see presented. In this case, it is quite clearly about the lack of representation of any female athlete among these top 100 earners. It does not need to be spelt out any more explicitly with captions or comments. It speaks for itself.

Sometimes a visualisation cannot speak for itself, but we can. I described earlier the imagined scenario of presenting a chart to an audience to get you to imagine what features you would point out and comment on. For explanatory visualisations you try to recreate this layer of insight directly within the work itself. However, if this was a real scenario you might use an exhibitory visualisation as your main prop and accompany this with your observations and gestures to provide an overall explanatory experience. I would define this as an *exhibitory* visualisation but with the understanding facilitated through an *explanatory* setting.

Furthermore, perhaps you are presenting a graphic as a figure within a written article or report. In and of itself the visualisation does not explain things in a stand-alone sense, but instead exists as a visual figure to reference from within and supplement the writing. The text therefore provides the explanatory narrative.

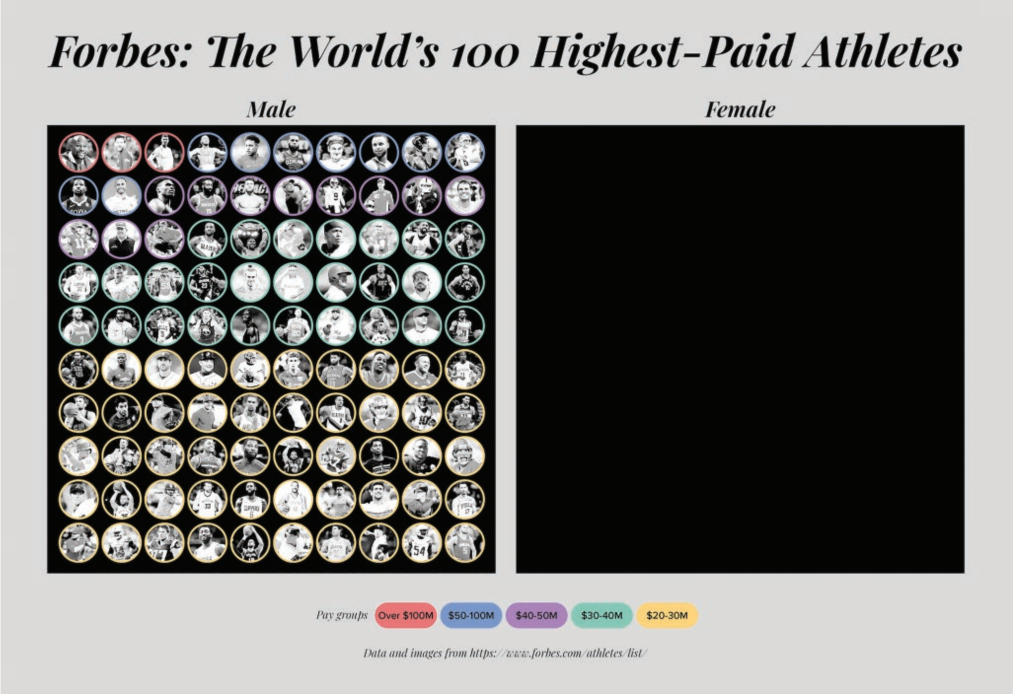


Figure 3.18 Forbes: The World's 100 Highest-paid Athletes, by Andy Kirk

Another common context for using an exhibitory visualisation might exist in the situation of producing a visual for stakeholders who have directly requested you to create something for them. They might not need to see anything other than the basic chart of data. They know what they are looking for and how to find it. The problem is that many visualisation projects mistakenly fall into the void of being exhibitory visualisations when they really needed to be more supportively explanatory or functionally exploratory.

Harnessing Ideas

The second aspect of establishing your project's vision offers an opportunity to harness imagination by capturing your initial, instinctive ideas. These are the earliest seeds of any thoughts you may have for what the eventual solution you are working towards might look like.

In *Thinking Fast and Slow*, author Daniel Kahneman describes two models of thought that control our thinking activities. He calls these System 1 and System 2 thinking: the former is responsible for our instinctive, intuitive and metaphorical thoughts; the latter is much more ponderous, by contrast, much slower, and requiring of more mental effort when being called upon. System 1 thinking is what you want to harness at this part of the first stage: what are the mental impressions that form quickly and automatically in your mind when you first think about the challenge you are facing?

You cannot switch off System 1 thoughts. Mental visualisations are what we instinctively 'see' in our mind's eye when we consider the subject or nature of a task. You will not be able to stop them happening when thinking about a problem. Rather than stifling your natural mental habits, this stage of the process presents the best possible opportunity to allow yourself space to begin imagining.

What colours do you see? Sometimes instinctive ideas are reflections of our culture or society, especially the connotations of colour usage. What shapes and patterns strike you as being semantically aligned with the subject? This can be useful not just to inspire but also possibly to obtain a glimpse into the similarly impulsive way the minds of your audience might connect with a subject when consuming the solution.

Think back to the example shown in Figure 3.9 about political 'buying power'. As a commonly recognisable metaphor of wealth, using Monopoly pieces was an entirely reasonable way to represent the data. Presenting this huge, imaginary pile on the lawn of the White House was symbolically congruent with the subject involved. The visualisation in Figure 3.19 concerns the wine industry, showing the top grape varieties grown. In the upper part of the graphic, the size of production for each grape variety is shown using a bubble chart, which creates a metaphorical representation of a bunch of grapes.

You can clearly see how this design might have been conceived from early ideas formed before the data was even collected and analysed. Not only is the representation consistent with the subject, but it also offers an immediately recognisable metaphor. Any viewer will make a seamless connection between subject and form.

To help unlock your imagination it is useful to be influenced and inspired by the world around you. Exposing your senses to different sources of influence can only help to broaden the range of solutions you might be able to conceive. Research the techniques that are being used across the visualisation field, look through books and see how others might have tackled similar subjects portraying similar types of data. Outside of the visualisation sphere, consider other forms of design or imagery: colours, patterns, shapes and metaphors from everyday life whose aesthetic qualities you just like. Start a scrapbook or project mood board that compiles the sources of inspiration you come across and helps you form ideas about the style, tone or essence of your project. They might not have immediate value for your current project but may materialise as useful for future work.

> 'I focus on structural exploration on one side and on the reality and the landscape of opportunities in the other … I try not to impose any early ideas of what the result will look like because that will emerge from the process. In a nutshell I first activate data curiosity, client curiosity, and then visual imagination in parallel with experimentation.' **Santiago Ortiz, Founder and Chief Data Officer at DrumWave, discussing the role – and timing – of forming ideas and mental concepts**

It is important to acknowledge the boundaries of this activity. Influence and inspiration are healthy: the desire to emulate what others have done is understandable. Plagiarism, copying and stealing uncredited ideas are wrong. There are ambiguities in any creative discipline about the boundaries between influence and plagiarism, and the worlds of visualisation and infographic design are not spared that challenge.

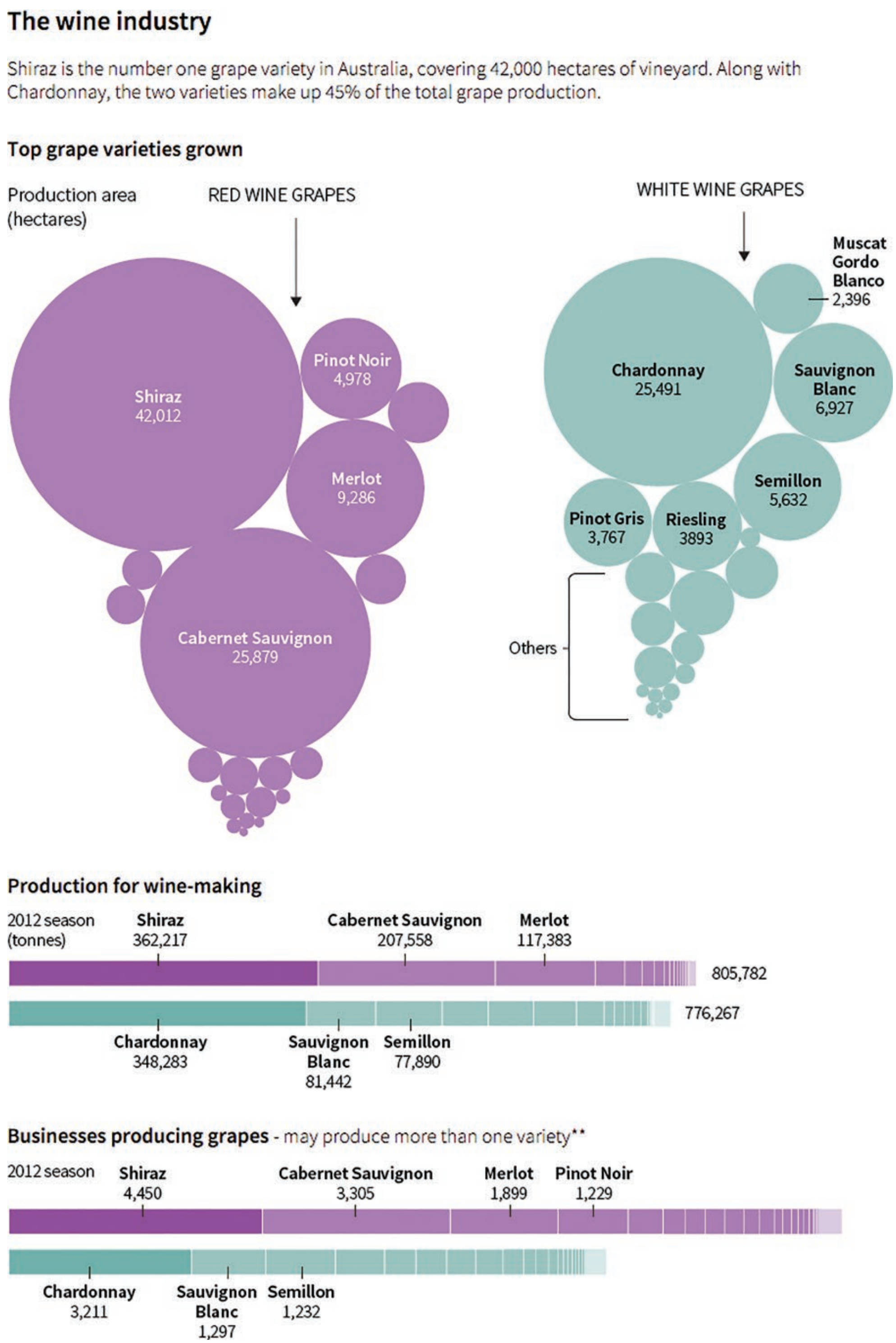


Figure 3.19 Grape Expectations, by S. Scarr, C. Chan and F. Foo (Reuters Graphics)

Being influenced by the research you do and the great work you see around the field is not stealing, but if you do incorporate in your work explicit ideas influenced by others, at the very least you should do the noble thing and credit the authors, or, even better, seek them out and ask them to grant you their approval. You do not have to credit William Playfair every time you use the bar chart, but there are certain unique visual devices that will unquestionably be deserving of attribution.

'Look at how other designers solve visual problems (but don't copy the look of their solutions). Look at art to see how great painters use space, and organise the elements of their pictures. Look back at the history of infographics. It's all been done before, and usually by hand! Draw something with a pencil (or pen … but NOT a computer!). Sketch often: The cat asleep. The view from the bus. The bus. Personally, I listen to music – mostly jazz – a lot.' **Nigel Holmes, Explanation Graphic Designer, on inspirations that feed his approach**

Sketching is also an important habit to develop. It does not require strong artistic talent, but it can prove to be a useful method for extracting ideas out of your mind and quickly capturing them in visual form. Figure 3.20 shows a montage of various sketches and sources of inspiration that influenced the design concept of a project visualising the spread of the #MeToo movement.

For some people, the most fluent and efficient way to 'sketch' is through their software application of choice rather than on paper. Regardless of the medium you use, sketching is useful when you are working with collaborators or for stakeholders as a means of discussing ideas, getting input and other's thoughts on the brief. I find it particularly helpful when trying to conceive innovative solutions to unusual or particularly complex challenges. It may be that my eventual solution looks nothing like my rudimentary sketches, but it gives me a way to cycle rapidly through iterations of concepts that may be worth exploring later on.

'It is easy to immerse yourself in a certain idea, but I think it is important to step back regularly and recognise that other people have different ways of interpreting things. I am very fortunate to work with people whom I greatly admire and who also see things from a different perspective. Their feedback is invaluable in the process.' **Jane Pong, Data Visualisation Designer**

There are limits to the value of ideas and the role they should be allowed to play. After all, data is your raw material, your ideas are not. It may be that your ideas are ultimately incompatible with the properties of the data you are working with, in which case you should just let go and move on.

Early sparks of inspiration in your thinking should be embraced, but do not be precious or stubborn. Always maintain an open mind and recognise that ideas have a limited role. This is why *harnessing* is the appropriate term used to describe this activity.

'I draw to freely explore possibilities. I draw to visually understand what I am thinking. I draw to evaluate my ideas and intuitions by seeing them coming to life on paper. I draw to help my mind think without limitations, without boundaries. The act of drawing, and the very fact we choose to stop and draw, demands focus and attention. I use drawing as my primary expression, as a sort of functional tool for capturing and exploring thoughts.' **Giorgia Lupi, Co-founder and Design Director at Accurat**



Figure 3.20 MeTooMentum, by Valentina D'Efilippo (design) and Lucia Kocincova (development)

Finally, it is worth noting the diplomatic prospect of taking on board other people's ideas. One of the greatest anxieties I face in my client work comes from working with stakeholders who are unequivocally and emphatically clear about what they think a solution should look like – from the very start.

Often your involvement in a project may arrive after these ideas have already been formed, during which time they have shaped the brief issued to you by the stakeholders ('Can you

make *this*, please?'). This is where your tactful but assured 'communicator' hat comes to the fore. The ideas presented to you may be reasonable and well intended, but it is your responsibility to lead the creation process. You can welcome input in the form of proposed concepts but, as with the limitations of your own ideas, there will be other factors with a greater influence: the nature of the data, the type of curiosities you are pursuing, the essence of the subject matter, and the nature of the audience, among many other things. These will be the factors that ultimately dictate whether any early vision of potential ideas ends up being of value.

Summary: Formulating Your Brief

This chapter commenced the opening stage of the design process concerned with initiating, defining and planning the requirements of your work.

Context

The first section looked at issues around context, specifically about the importance of defining the motivating curiosity and identifying all the circumstances that will shape your project. These included factors such as follows.

People

- Stakeholders: Who is the ultimate customer? Who are the influencers, interferers, subject matter experts (SMEs)?
- Audience: What is their knowledge (informed or 'layperson')? Receptive or indifferent?
- Visualiser(s): What skills/knowledge are possessed? Individual or team?

Constraints

- Timescales: When is it due? When can you start? Milestones? Available duration?
- Pressures: Financial? Political? Cultural? Environmental?
- Design: Style restrictions (colour, type, logo), size?
- Technological: What software, hardware, infrastructure exist? Platform compatibility?

Deliverables

- Setting: Rapid or prolonged? Consumed remotely or live?
- Medium: What is the intended output format?
- Quantity: How many outputs are being produced?
- Frequency: One-off project or a regular/repeated task?

Vision

The second section considered the vision of your work, firstly looking at its core purpose. What is it for? What are you trying to accomplish? Depending on your defined purpose, you will need to pursue the right balance in the tone and experience through which understanding will be facilitated:

- Tone: The distinction between 'reading' and 'feeling' data.
- Experience: The difference between 'explanatory', 'exhibitory' and 'exploratory' visualisations.

Finally, you learnt about the value and limitations of harnessing ideas. What mental images, shapes, forms and keywords instinctively come to mind when thinking about the subject matter of this challenge? What influence and inspiration can you source from elsewhere that might start to shape your thinking?

General tips and tactics

- Not all circumstantial factors can be defined, nor will they be stable throughout. Certain things may change in definition, some undefined things will emerge, some defined things will need to be reconsidered, some things are just always constraint-free.
- Notes are so important to keep about any thoughts you have had that express the nature of your curiosity, articulation of purpose, any assumptions, things you know and do not know, where you might need to get data from, who the experts are, questions, things to do, issues/problems, wish lists, etc.
- Keep a 'scrapbook' (digital bookmarks, print clippings) of anything and everything that inspire and influence you – not just data visualisations. Log your ideas and inspire yourself.
- This stage is about ambition management and it will be to your benefit if you treat it with the thoroughness it needs. The impact of any corners being cut here will be amplified later on.

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

4

Working With Data

In Chapter 3, the design process was initiated, with early attention paid to defining matters of context and vision. The discussion about context looked at identifying the source curiosity and the circumstances that would influence the conditions of your work. Vision was a more forward-facing glance towards the purpose of your work, thinking about how you might accomplish the type of understanding you are seeking to facilitate for your audience. We closed the chapter by looking at the value of harnessing ideas, through sketching and research.

In this second chapter, your thinking will switch to the practical mechanics of working with data. In this chapter you will be respecting its role as the critical raw material of this process, learning how to nurture its potential but equally being prepared for it to frustrate you, on occasion. You will work through four distinct activities to develop a close acquaintance with it, as follows:

- Data acquisition: Sourcing and gathering the raw material.
- Data examination: Familiarising yourself with the key physical properties and condition of your data.
- Data transformation: Refining your data through modification and consolidation.
- Data exploration: Using exploratory analysis and research techniques to discover insights.

I often encounter people who declare their love for data. Data does indeed have the capacity to earn and merit love, but I personally do not love it all that much. Data always demands so much attention yet consistently seems to conspire against you. You do not need to love data but, equally, you should not fear data – you should just respect it.

Some readers might feel confident working with data but might not have much direct experience when working with it in the context of a visualisation challenge. For those readers I want to provide you with a strong appreciation of the influence data has on your editorial and design thinking and dispel any sense that this is especially complicated. For those with more experience and confidence with this topic, this chapter might help to reinforce some of the ways of thinking about the impact of your data on your visualisation project.

4.1 Step 1: Data Acquisition

The first step when working with data in a visualisation project is to get the data. There are several distinct origins and methods involved in acquiring data. Some are characterised by your having to do most of the work yourself, others involve people making it available for you to access in different ways. In each of these cases you need to be assured about the reliability of the data you are gathering, whether it is you or others who have curated it. As discussed in Chapter 2 when describing the importance of 'trustworthiness', there may be collection issues creating inaccuracies and biases that can affect the quality of your data at source. You need to be discerning in the degree of trust you place in it, at least to begin with, until you have a chance to examine it more closely.

Supplied: The simplest method for acquiring data involves getting it from somebody else. In projects where you have been commissioned by a stakeholder (manager, client or colleague), you will often be issued with the data you need. The extent of your efforts may therefore just involve saving an attachment from an email. You should, however, still undertake as much background research as possible about the original source and collection method used to form the data you have been given.

System download: When working in organisations, there will inevitably be many occasions where the data will come from internal reporting tools or exports from corporate systems. There are many organisations offering publicly accessible data to download through the Web, sometimes through interfaces that let interested users construct detailed queries and download structured data customised to their need. There is an increasing marketplace for data.

Web scraping: This involves using special programs to extract structured and unstructured items of data published on web pages. For example, you may wish to extract information from a hotel chain, product details from the IKEA website or data about the history of the Winter Olympics on Wikipedia. Depending on the tools used, you can often set routines in motion to extract data across multiple pages of a site based on the connected links that exist within it. This is known as *web crawling*. An important consideration to bear in mind with any web scraping or crawling activity concerns rules of access and the legalities of extracting the data held on certain sites. Always check – and respect – the terms of use before undertaking this.

APIs: Certain specialist websites or services offer an API (Application Programming Interface) to enable people to access streams of data. A popular example would be the access provided to real-time data on topics like air quality, traffic disruptions and London Underground passenger levels provided by Transport for London (TfL), which encourages people to develop bespoke software applications. Many commercial services now offer extensive sources of curated and customised data that would otherwise be very complex to gather or difficult to obtain. An example might include large, customised extracts from social media platforms like Twitter based on specific keyword criteria.

Primary collection: If the data you need does not exist in digital form, you might need to consider gathering primary data. This is where you collect observations or capture measurements about bespoke phenomena specific to your needs. These might include:

- Transcribing a political speech from unstructured video or audio recordings to explore patterns of sentiment and/or rhetoric.
- Designing a participant questionnaire to collect relevant data about a research study.
- Using measurement devices to track fitness activity or health information over a period of time.

This type of data gathering activity can be expensive in time and cost. The benefit is that you will be able to control carefully the collection of the data to ensure its value is optimised for your needs.

'Don't underestimate the importance of domain expertise. At the Office for National Statistics (ONS), I was lucky in that I was very often working with the people who created the data – obviously, not everyone will have that luxury. But most credible data producers will now produce something to accompany the data they publish and help users interpret it – make sure you read it, as it will often include key findings as well as notes on reliability and limitations of the data.' **Alan Smith OBE, Data Visualisation Editor, *Financial Times***

Data foraging: If the data you need does not exist in a convenient single form or location, you may need to *forage* for it. This usually involves manually sourcing relatively small amounts of disparate or dispersed data values. For example, suppose you wanted to compare the lifetime costs associated with a range of different mobile phones to help you decide which model or tariff to go with. You would find the information on the Web, locate the specific data items and values you need, collect them in a spreadsheet, and then repeat for each model or tariff to build up your table of comparable data. Sometimes data foraging involves extracting data from documents, such as pdf files. There are tools that assist in accelerating this task, enabling you seamlessly to extract tables of data and convert them into more usable Excel or CSV (Comma-Separated Values) formats.

4.2 Step 2: Data Examination

'Data inspires me. I always open the data in its native format and look at the raw data just to get the lay of the land. It's much like looking at a map to begin a journey.' **Kim Rees, Head of Data Experience Design at Capital One**

Once you have acquired your data – whether this is all of it or just a starting point – the second step is to examine it thoroughly. Before you choose what meal to cook, you need to know what ingredients you have, how much and in what condition. The same applies to data. Before you can contemplate any design thinking you first need to familiarise yourself

fully with the physical characteristics and state of your data. Examining your data specifically involves learning about the types of data you have, the size and range of values held, and its condition.

Data Types

Before we look specifically at a classification for different types of data, we first need to establish the difference between types of tabulation. A *normalised* form of tabulated data offers the most detailed, granular form of data, organised by *variables* (columns or fields) and *items* (rows or records). The table in Figure 4.1 is a simple, small-scale example of a normalised dataset. Each column in the table represents a different variable describing movies in each film series.

Figure 4.1
Example of a Normalised Dataset

Film Series	Movie Title	Rotten Tomatoes Score	Year Released
Harry Potter/J K Rowling	Harry Potter and the Sorcerer's Stone	80%	2001
Harry Potter/J K Rowling	Harry Potter and the Chamber of Secrets	82%	2002
Harry Potter/J K Rowling	Harry Potter and the Prisoner of Azkaban	90%	2004
Harry Potter/J K Rowling	Harry Potter and the Goblet of Fire	88%	2005
Harry Potter/J K Rowling	Harry Potter and the Order of the Phoenix	77%	2007
Harry Potter/J K Rowling	Harry Potter and the Half-Blood Prince	84%	2009
Harry Potter/J K Rowling	Harry Potter and the Deathly Hallows - Part 1	78%	2010
Harry Potter/J K Rowling	Harry Potter and the Deathly Hallows - Part 2	96%	2011
Harry Potter/J K Rowling	Fantastic Beasts and Where to Find Them	74%	2016
Star Wars	Episode IV: A New Hope	93%	1977
Star Wars	Episode V: The Empire Strikes Back	95%	1980
Star Wars	Episode VI: Return of the Jedi	80%	1983
Star Wars	Episode I: The Phantom Menace	55%	1999
Star Wars	Episode II: Attack of the Clones 3D	66%	2002
Star Wars	Episode III: Revenge of the Sith 3D	79%	2005
Star Wars	Episode VII: The Force Awakens	93%	2015
Star Wars	Rogue One	84%	2016
Star Wars	Episode VIII: The Last Jedi	91%	2017
Star Wars	Solo: A Star Wars Story	70%	2018
Tolkien/Middle Earth	The Lord of the Rings: The Fellowship of the Ring	91%	2001
Tolkien/Middle Earth	The Lord of the Rings: The Two Towers	95%	2002
Tolkien/Middle Earth	The Lord of the Rings: The Return of the King	93%	2003
Tolkien/Middle Earth	The Hobbit: An Unexpected Journey	64%	2012
Tolkien/Middle Earth	The Hobbit: The Desolation of Smaug	74%	2013
Tolkien/Middle Earth	The Hobbit: The Battle of the Five Armies	59%	2014
X-Men	X-Men	82%	2000
X-Men	X2: X-Men United	85%	2003
X-Men	X-Men: The Last Stand	58%	2006
X-Men	X-Men Origins - Wolverine	37%	2009
X-Men	X-Men: First Class	86%	2011
X-Men	The Wolverine	70%	2013
X-Men	X-Men: Days of Future Past	90%	2014
X-Men	Deadpool	84%	2016
X-Men	X-Men: Apocalypse	48%	2016
X-Men	Logan	93%	2017
X-Men	Deadpool 2	83%	2018

In contrast, *cross-tabulated* datasets present a summarised form of normalised data, displaying values that are the result of statistical operations such as grouped totals, maximums

and minimums. If normalised data is sometimes colloquially described as being 'tall and thin', cross-tabulated data is 'short and fat'. In Figure 4.2, you will see a cross-tabulated form of the data shown in the normalised table in Figure 4.1.

Film Series	Number of Movies	Best Rotten Tomatoes Score	Worst Rotten Tomatoes Score	Year of First Release
Harry Potter/J K Rowling	9	96%	74%	2001
Star Wars	10	95%	55%	1977
Tolkien/Middle Earth	6	95%	59%	2001
X-Men	11	93%	37%	2000

Figure 4.2 Example of a Cross-tabulated Dataset

If you are working with data that exists in cross-tabulated form and you do not have access to the underlying normalised data, the avenues of potential analysis will be reduced in scope. As you can see in Figure 4.2, you have a far reduced set of data items and values to work with. There is no granularity, no sense of the distribution and variety of values that lie beneath these statistical aggregates.

The type of data tabulation is also influential when conducting analysis and generating charts. Certain tools need data to be shaped in specific ways. For example, charting in Excel is usually performed by linking a selected chart template to a range of cells that are usually organised in cross-tabulated form. Working with a tool like Tableau involves connecting to normalised data and constructing a chart from the ground up. I find it always preferable to work with normalised data, where you have far more detail that you can then choose to aggregate should you wish. The key is that you have the choice.

For the purpose of this chapter, we will principally look at working with data in normalised form.

Next you need to develop a thorough understanding of your *data types*. Also commonly referred to as *levels of data* or *scales of measurement*, data types define the nature of the values held under each variable and about each item in your dataset. The different types of data you might have will have a major influence over several key aspects of the design process, such as:

- determining the type of statistical analysis methods you can use;
- shaping the editorial perspectives you will pursue;
- filtering the specific chart types you can or cannot use;
- informing the appropriateness of your colour associations; and
- guiding composition decisions on size, placement and layout.

In simple terms, data types are distinguished by being either qualitative or quantitative in nature. Below this high-level view there are more nuanced but crucial distinctions that need to be understood.

The most useful taxonomy I have found to distinguish types of data, in the context of developing a visualisation, comes from psychologist researcher Stanley Stevens. He devised the NOIR classification which represents: Nominal, Ordinal, Interval and Ratio. The order of these distinct types is deliberate, as each subsequent level of measurement embodies a certain

> 'Absorb the data. Read it, re-read it, read it backwards and understand the lyrical and human-centred contribution.' **Kate McLean, Smellscape Mapper and Senior Lecturer Graphic Design**

increase in precision. This is a particularly relative approach to working with data in the context of social research. I find it useful to extend the acronym, adding a leading 'T' – for Textual – to reflect the contemporary experiences of working with a greater variety of qualitative data.

Textual data is qualitative and characteristically 'human', usually existing in unstructured form like passages of text or sections of a report. Examples of textual data might include:

- Responses to 'Any other comments?' in a questionnaire.
- A newspaper write-up about a football match.
- The abstract for an academic research article.
- The description of a product on Amazon.
- The transcript of a speech given by a politician.

When working with textual data you will typically need to transform it to extract certain properties and relational characteristics in some way, such as counting the frequency of certain keywords or using natural language processing techniques to derive sentiment classifications.

Nominal data is another form of qualitative data and the second distinct data type. Nominal data exists as categories, which offer means of separating different values and grouping similar values together. Examples of nominal data might include:

- The gender of a survey participant.
- The meals available on a restaurant menu.
- The name of your country of birth.
- The genre of a movie.
- The sport events in the Olympics.

Nominal data does not exclusively mean text-based data; nominal values can be numeric. For example, a student ID number is a categorical device used to uniquely identify all student records. The shirt number of a footballer is a way of helping teammates, spectators and officials recognise each unique player. You might find measurements of gender captured as 1 (male), 2 (female) and 3 (other), but these numeric values should not be considered quantitative values – adding '1' to '2' does not equal '3', for gender.

Another characteristic of nominal data is the potential for a hierarchical relationship to exist between two or more major and sub-categorical variables. For example, a major category holding details of 'Country' and a sub-category holding 'Airport'; or a major category holding details of 'Industry' and a sub-category holding details of 'Company Names'. Recognising this type of relationship will become important when deciding how to portray your data using certain chart types.

Ordinal data is the third qualitative data type. Unlike nominal data, ordinal data is characterised by their being some notion of order in the relationship between different categorical values. Examples of nominal data might include:

- The options available to answer a survey question based on the extent to which you agree or disagree with a statement.
- General temperature observations from *very hot* to *very cold*.
- The size of T-shirts from XS to XXL.
- The rank of a police officer.
- The gold, silver and bronze medal categories at the Olympics.

Recognising a categorical variable as being ordinal rather than nominal in nature will be particularly relevant when you make decisions about classifying values using different colour scales.

Interval data is a quantitative measurement defined by difference on a scale but *not* by relative scale. Examples of interval data might include:

- The body mass index for measuring obesity.
- The forecasted temperature in °C.
- The latitude and longitude coordinates of a given location.

The principal characteristic of interval data is that the absolute difference between two values is meaningful, but any arithmetic operation, such as multiplication, is not. For example, the absolute difference between 15°C and 20°C is the same difference as between 5°C and 10°C. However, the relative difference between 5°C and 10°C is not the same as the difference between 10°C and 20°C (where in both cases you multiply the lower value by two or increase by 100%). This is because the zero state of an interval scale is not a true zero value, it is just an established scale position. A temperature reading of 0°C does not mean there is no temperature; it is a quantitative scale for measuring relative temperature.

Ratio data is the second quantitative type of data and the one you are most likely to encounter in most visualisation project situations. Examples of ratio data might include:

- The age of a survey participant in years.
- The forecasted amount of rainfall in millimetres.
- The estimated budget for a research grant proposal in GBP (£).
- The number of sales of a book on Amazon.
- The distance of the winning long jump at the 2016 Olympics in metres.

Ratio data values are numeric measurements with significant properties of both difference *and* scale. The absolute difference in age between a 10- and 20-year-old is the same as the difference between a 40- and 50-year-old. The relative difference between a 10- and a 20-year-old is the same as the difference between a 40- and an 80-year-old ('twice as old'). Unlike interval data, a zero value for a ratio variable is a true zero, meaning there is no amount.

There are other important data-type distinctions. One key distinction is between values that are *discrete* or *continuous*. This distinction is particularly influential in how you might visually represent the relationship between such values. Discrete data is associated with all classifying measurements that have no 'in-between' state. This applies to all qualitative data types and any quantitative values for which only a whole is possible. Examples include the heads or tails outcome of a coin toss, the days of the week and the number of seats in a cinema.

In contrast, continuous measurements can hold the value of an in-between state and indeed any value between the natural upper and lower limits, if such fine degrees of measurement detail are possible. Think of them as moment-in-time measurements, with examples including height, weight and temperature.

There are some data-type classifications that are hard to define on the TNOIR scale, due to special or varied characteristics that are not universally compatible with this taxonomy. One such example would be time-based data, which can shift across the TNOIR classification depending on the format and purpose of using this data (e.g. for grouping, for labelling or for calculations of duration).

Whereas most quantitative measurements you will deal with are based on a linear scale, there are exceptions. Variables about the strength of sound (decibels) and magnitude of earthquakes (Richter) are actually based on a logarithmic scale. An earthquake with a magnitude of 4.0 on the Richter scale is 100 times bigger and 1000 times stronger (based on the amount of energy released) than an earthquake of magnitude 2.0. Logarithmic values, as well as other mathematically derived types of data, are often still considered as ratio variables but are distinguished as being non-linear scaled variables.

Data Size: Amount and Range

Once you have established an understanding of the different types of data, you can switch your examination towards the shape and size of this data, looking at the quantitative attributes across all variables and for all items. The main questions you will ask of your data include:

- For quantitative variables (interval or ratio), what is the lowest and the highest value in each case?
- In what format are the numeric values presented (i.e. how many decimal places or comma-formatted)?
- For a categorical variable (nominal or ordinal), how many different values are held?
- If you have textual data, what is the maximum and minimum character length or word count?

Statistical methods will assist in describing further physical characteristics. Here are some analyses you might find useful at this stage. These are not the *only* methods you will ever need to use, but it is likely they will be the most common.

- *Frequency distribution:* Applied to quantitative values to learn about the shape of the distribution of values.
- *Measurements of central tendency:* These describe the summary attributes of a group of quantitative values, including: the mean (the average value); the median (the middle value if all quantities are arranged from smallest to largest); the mode (the most common value).
- *Frequency counts:* Applied to categorical values to understand the frequency of different instances.
- *Measurements of spread:* These are used to describe the dispersion of values above and below the mean:
  - Maximum, minimum and range: the highest and lowest and magnitude of spread of values.
  - Percentiles: the value below which *x*% of values fall (e.g. the 20th percentile is the value below which 20% of all quantitative values fall).
  - Standard deviation: a calculated measure used to determine how spread out a series of quantitative values are.

As mentioned at the end of the previous chapter, it is worth repeating that your ideas may stimulate certain design thinking, but the shape and size of your data will drive it. The quantitative characteristics of your data will have a strong bearing on what may or may not qualify as a suitable design solution. For example, look at the shape of data in the 'Better Life Index' project that you saw earlier. As you can see in Figure 4.3, the analysis of the quality of life covers 38 OECD member states and uses a flower structure for each country comprising 11 petals. Each petal represents a different quality of life indicator.

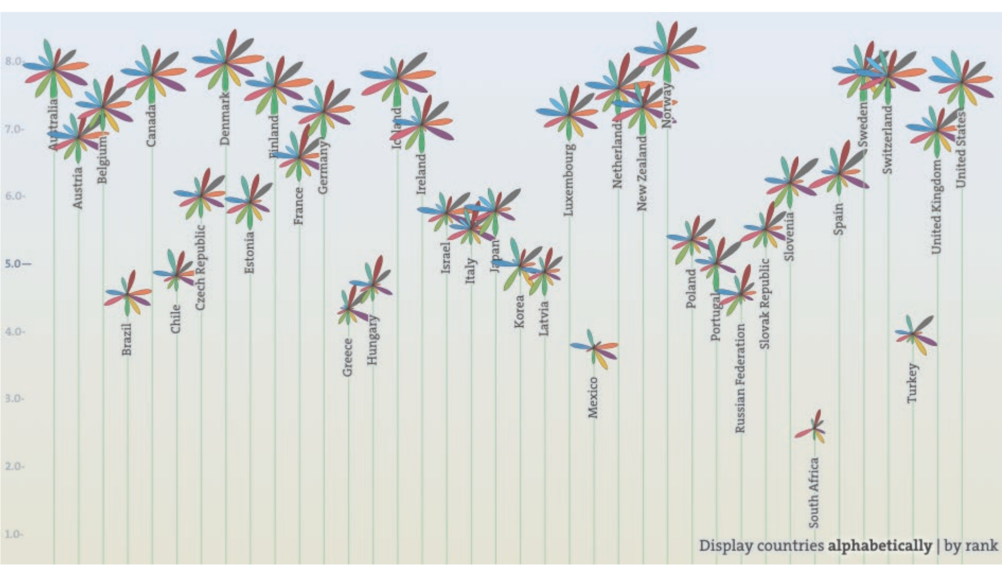


Figure 4.3 OECD Better Life Index, by Moritz Stefaner and Dominikus Baur, Raureif GmbH

What if there were 25 indicators? Or just 3? What if the analysis was expanded to cover 150 countries? Would the idea of using the flowers as a metaphor for quality of life still work for such different dimensions of data? Arguably not.

> 'My design approach requires that I immerse myself deeply in the problem domain and available data very early in the project, to get a feel for the unique characteristics of the data, its "texture" and the affordances it brings. It is very important that the results from these explorations, which I also discuss in detail with my clients, can influence the basic concept and main direction of the project. To put it in Hans Rosling's words, you need to "let the data set change your mind set".' **Moritz Stefaner, Truth & Beauty Operator**

Another relevant example of the impact of the shape of data can be demonstrated in the 'Spotlight on profitability' graphic in Figure 4.4. Although there are several movies that slightly exceed the maximum scale value of $1000 million, there is one movie, *Avatar*, that emphatically bursts through the spatial ceiling of the chart. As an isolated but extreme outlier, in this case, it was accommodated using a bespoke approach. As you can see, the towering shape of this movie trespasses above into the space afforded by two empty rows. This solution emphasises the value's exceptional quality.

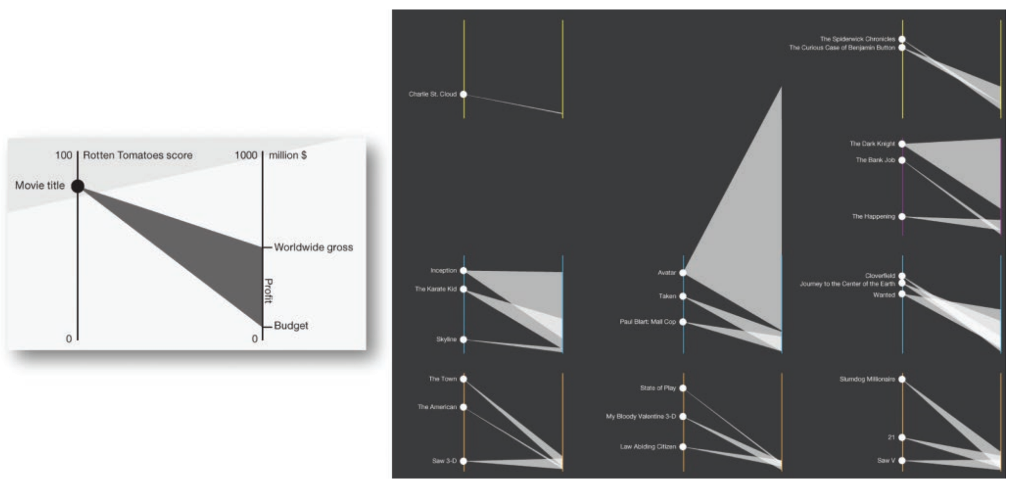


Figure 4.4 Spotlight on Profitability, by Krisztina Szücs

How do you elegantly handle quantitative measures that have hugely varied value ranges? Accommodating the full range of your data values into a single chart scale can have a distorting impact on the space it occupies. Normally, you would not have the luxury of being able to apply a customised approach to handling such diverse data values. At this stage, we are not so concerned about working out a design solution, rather the general point is to become aware of the existence of the 'Avatars' in your data that will eventually have an impact on your design thinking.

Data Condition: Quality and Representativeness

Undiscovered and unresolved issues around the quality of your data will undermine the trust in and the accuracy of your work. You will need to discover and address these issues during this stage of the process. Features to look out for may include:

- Missing values: Are empty cells assumed to be of no value (zero/nothing) or no measurement (n/a, null)? This is a subtle but important difference.
- Erroneous values: Typos and any values that clearly look out of place (such as a gender value in an 'age' column).
- Inconsistencies: Capitalisation, units of measurement, value formatting.
- Duplicate records.
- Expired values: Values that might have elapsed in their current relevance or accuracy, like someone's age or any statistic that would be expected to have subsequently changed.
- Uncommon system characters or line breaks.
- Leading or trailing spaces: A subtle but particularly evil issue!
- Date issues around format (dd/mm/yy or mm/dd/yy) and basis (systems like Excel's base dates on daily counts since 1 January 1900, but not all do that).

An extension of examining the condition of your data is to consider how representative it is. This is, in part, about appreciating what is missing, not just in value terms, but more in relation to the items of data you have and do not have about your subject matter. You need to be healthily sceptical about your data, seeking constant reassurance of its quality and condition, so you can be confident that what you are presenting is legitimate. Inaccuracies in judging what your data truly represents can have an even greater impact on trust than the damage caused by individual elements of missing or inaccurate data. The questions you need to ask of your data are:

- Does it represent genuine observations about a given phenomenon or is it influenced by the limitations of a collection method?
- Does your data reflect the entirety of a particular phenomenon, a recognised sample, or maybe even an obstructed view caused by hidden limitations in the availability of data about that phenomenon?

It is simply not reasonable to expect always to have access to the entirety of data about your subject matter. Most projects you will work with will resemble a sample of a population. This is not an obstacle to progressing with a visualisation, it is about caution rather than cessation. You must be clear about the basis on which your sample is formed and how you might faithfully represent and communicate this to your audience. You might even exploit the gap that exists between the data you have and all the data that could exist about the phenomena to shine a light on what is missing. Make that your key focus.

4.3 Step 3: Data Transformation

Once you complete your examination of your data you will have a good idea about what actions may be needed to transform your data. This is to prepare it for the analysis and charting steps you will soon move on to. What do you need to do to get it into shape and fit for purpose?

It is worth noting that transforming your data is a prime example of a step in this process that may start now but will likely continue right the way through to the latter stages of your design thinking. As you reach that stage you will often encounter the need to tweak further the shape and size of your data.

In accordance with the desire for trustworthy design, any modifications or enhancements you apply to your data need to be noted and potentially explained to your audience. You must be in a position to share the thinking behind any significant assumptions, calculations and conversions you have made.

There are three different types of potential activity involved in transforming your data: cleaning, creating and consolidating.

> 'When I first started learning about visualisation, I naively assumed that datasets arrived at your doorstep ready to roll. Begrudgingly I accepted that before you can plot or graph anything, you have to find the data, understand it, evaluate it, clean it, and perhaps restructure it.'
> **Marcia Gray, Graphic Designer**

Cleaning: I have already discussed the importance of data quality. There is no need to revisit the list of potential issues to look out for, but the point is that now is the time to begin to address these.

There is no single approach for how best to conduct data cleaning. Some issues can be resolved through a straightforward 'find and replace' (or remove) operation. Other tasks might be much more intricate, requiring manual intervention, often in combination with inspection features like *sorting* or *filtering*, to find, isolate and modify any problem values.

A further part of cleaning your data involves eliminating what you do not need. Any variable or items of data that serve no ongoing value will take up space and attention. My tactic is usually to gather as much data as I can initially and then remove or at least archive it later to help reduce the clutter.

Remember, also, to keep backups. Before you undertake any transformation, make a copy of your dataset. After each major iteration, save further copies. It can be useful to preserve your original unaltered data so you can easily return to that state should you ever need to.

Creating: The most substantial transformation work often comes in the form of creating new data from existing values. This task is something I refer to as the hidden cleverness, where you expand your data to form new calculations and derive new groupings or any other mathematical treatments. In doing so you broaden the range of analytical options open for you to explore. There are unlimited different approaches to doing this depending on the data you have and what you need from it, though they might at least include:

- Creating percentage calculations based on existing quantities.
- Creating a calculation to establish a rolling 12 monthly total.

- Using 'start date' and 'end date' values to calculate the duration in days.
- Converting absolute quantities associated with different geographic locations into 'per capita' values based on population numbers in each.
- Using logic-based formulae to create new categorical values out of quantities, such as checking if an age value is under 18, in which case the 'Age group' value would be 'Child', otherwise 'Adult'.

I mentioned earlier in this chapter how sometimes your data does not exist in a tabulated form, but instead in an unstructured, qualitative document. In these cases you may choose to derive reasonable categorical or quantitative values from the original form. For example, when performing categorical transformations from textual data, you might seek to:

- Identify keywords or summary themes from text and convert these into categorical classifications.
- Identify and flag up instances of certain cases existing or otherwise (e.g. X is mentioned in this passage).
- Identify and flag up the existence of certain relationships (e.g. A and B were both mentioned in the same passage, C was always mentioned before D).
- Use natural language processing techniques to determine sentiments, to identify specific word types (nouns, verbs, adjectives) or sentence structures (around clauses and punctuation marks).
- With URLs, isolate and extract the different components of website address and sub-folder locations.

Similarly, for quantitative transformations of textual data, here are some common approaches:

- Calculate the frequency of certain words being used.
- Analyse the attributes of text, such as total word count, physical length, potential reading duration.
- Count the number of sentences or paragraphs, derived from the frequency of different punctuation marks.
- Position the temporal location of certain words/phrases in relation to other words/phrases or compared with the whole (e.g. X was mentioned at 1m 51s).
- Position the spatial location of certain words/phrases in relation to other words/phrases or compared with the whole.

Figure 4.5 presents a graphic showing an analysis of the profanities used by CEOs from a review of recorded conference calls over a period of time. This work demonstrates two ways of utilising textual data in visualisation. Firstly, the visualiser has extracted categorical classifications and quantitative measurements to show the trends in usage over time and to compare the frequency of different, certain swear words being used. Secondly, we see annotated captions lower down the page which preserve the original qualitative form of the textual data, without any transformation applied. This provides the viewer with some examples of the context of a swear word shown within a sample passage of the original transcript.

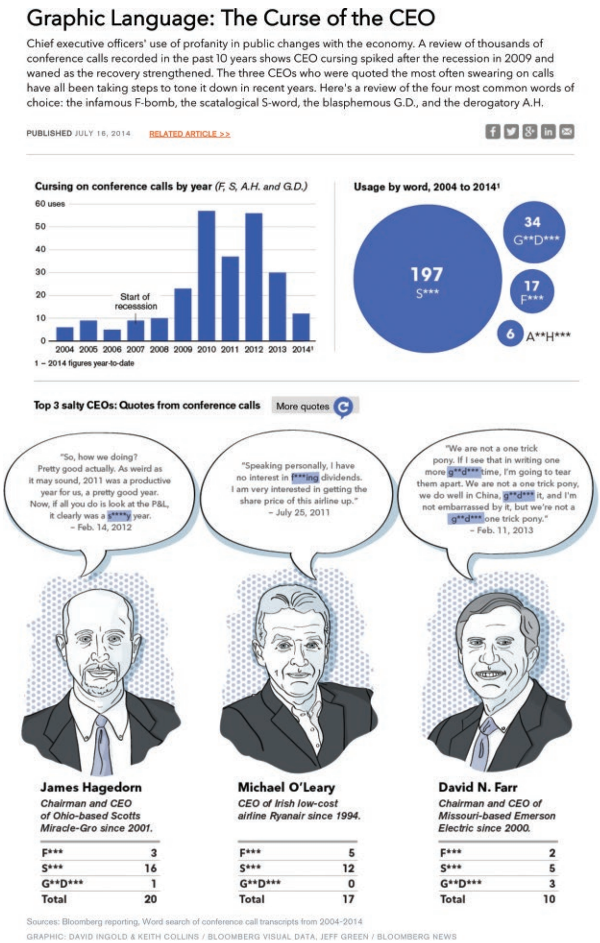


Figure 4.5 Graphic Language: The Curse of the CEO, by David Ingold and Keith Collins (Bloomberg Visual Data) and Jeff Green *(Bloomberg News)*

> 'Although all our projects are very much data driven, visualisation is only part of the products and solutions we create. This day and age provides us with amazing opportunities to combine video, animation, visualisation, sound and interactivity. Why not make full use of this? Judging whether to include something or not is all about editing: asking "is it really necessary?". There is always an aspect of gut feel or instinct mixed with continuous doubt that drives me in these cases.'
> **Thomas Clever, Co-founder CLEVER°FRANKE, a data-driven experiences studio**

Handling textual data will always create more work and you will need to judge the reward vs effort of such activity: how much effort will I need to expend in order to extract usable, valuable content from the text? Some of the approaches you might use will be quite straightforward to undertake, but others are more complicated and require sophisticated tools to assist with more algorithmic techniques.

A further transformation task involves converting the layout and format of your data. Formatting data for its appearance in printing

is not usually compatible with how we need it arranging when analysing or visualising it. For example, you might need to go through a spreadsheet and 'unmerge' any cell values that are formatted across several table columns. Sometimes you might encounter visual formatting like background shading or the colouring of a font to represent a key status. This might be useful when reading the table, but few tools will be able to 'see' these attributes – they will need to exist as actual values.

Consolidating: There will be occasions where you may seek to source and introduce additional data to expand (more variables) or append (more items) your data further in order to enhance its analytical potential:

- *Expand:* This is where you want to broaden the values of data you have to work with. For example, if you have location data at the level of detail of country, you might want to group and aggregate your analysis to a higher level. You could therefore source continent groupings, so you can then create this hierarchical relationship, giving you the option to analyse at both levels.
- *Append:* This might occur if your original dataset is no longer representative of the most up-to-date state and newer data items are available for you to access. If you are doing an analysis about movies, as soon as another week elapses new movies will be released, and your existing data will no longer have the most current items.

You may also use this moment in the process to start thinking about sourcing other assets that might enhance your data representation options later on. Perhaps the elegance of your work will be improved through possible access to photo-imagery, written anecdotes, video clips or physical media?

Even though it will be a while until we reach the design thinking stage, it is useful to start thinking about this as early as possible in case the collection of these additional assets requires significant time and effort. It might also reveal to you any particular obstacles involved in obtaining permissions for use or blockages to sourcing high-quality media. If you know you are going to have to do this asset gathering, do not leave it too late – reduce the possibility of such stresses by acting early.

4.4 Step 4: Data Exploration

Widening the Viewpoint

The examination step was about forming a deep acquaintance with the physical properties of your data. In doing this you will have found reasons and ways to enhance the data by transforming it. Next, you ideally need to build in some time to interrogate your data further to give yourself every opportunity of discovering the potential insights and qualities of understanding your data may provide. This is where we embark on visual exploration, an activity that is especially important when you are working with big datasets and/or datasets that are unfamiliar to you.

> 'After the data exploration phase you may come to the conclusion that the data does not support the goal of the project. The thing is: data is leading in a data visualisation project – you cannot make up some data just to comply with your initial ideas. So, you need to have some kind of an open mind and "listen to what the data has to say" and learn what its potential is for a visualisation. Sometimes this means that a project has to stop if there is too much of a mismatch between the goal of the project and the available data. In other cases this may mean that the goal needs to be adjusted and the project can continue.' **Jan Willem Tulp, Data Experience Designer**

Undertaking data exploration involves the use of both statistical and visual techniques to help you go beyond just *looking* at data and to begin to start *seeing* it. What answers to your overriding curiosity can you find? What other enlightening features of your data can you unearth? Sometimes, it might present new discoveries that will motivate you to pursue different avenues of enquiry.

Overall, you are trying to widen your viewpoint and be truly acquainted with the full potential of what your data is offering you. As I have emphasised about the benefit of this whole design process, to make the best decisions you first need to be aware of all the options. Data exploration is about broadening your awareness of the potentially interesting things you *could* show your audience about your subject. Making choices about ultimately which ones you *will* pursue comes next.

To frame this discussion I find it useful to refer to the transcript of a news briefing given by the then US Secretary of Defense, Donald Rumsfeld, in February 2002. This was his infamous 'known knowns' statement:

> Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones.

There has been much said about the apparent lack of elegance in the language used and criticism of the muddled meaning in this passage, but I disagree. I think it is probably the most concise way he could have articulated this, at least in written or verbal form. The essence of this statement, as presented in visual form in Figure 4.6, was to distinguish between an *awareness* of what is knowable about a subject from the status of *acquiring* this knowledge. When thinking about the role of data exploration, there is much we can gain from this concept.

The *known knowns* are aspects of understanding present in your data about your subject that you are aware of: you know about these things. When working with familiar data it can be tempting to consider the known knowns as being the only relevant perspectives to build your analysis around. Indeed, your knowledge of the subject may have influenced the nature of the curiosity you are pursuing.

However, what you know to be interesting about a subject may only represent a quite narrow viewpoint. It is therefore important to be willing to seek other interesting qualities of your

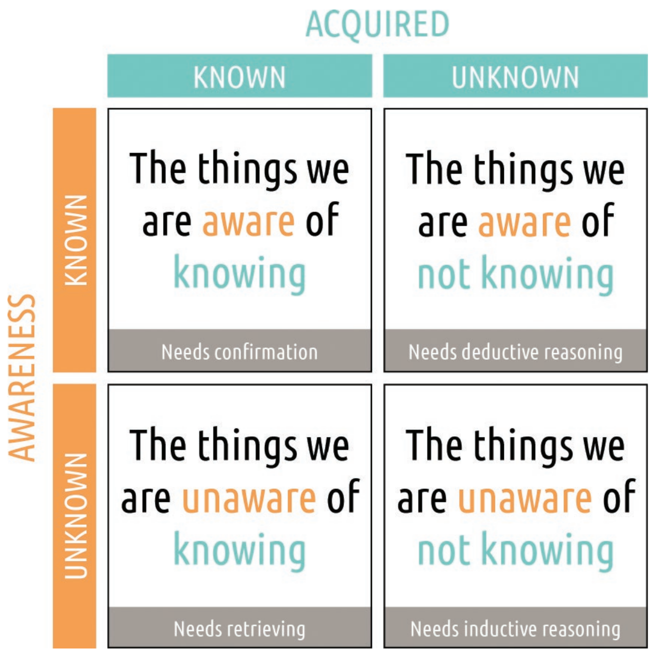


Figure 4.6 Making Sense of the *Known Knowns*

data, especially if it or the subject is unfamiliar. Indeed, your known knowns may be entirely absent if you know nothing about a subject and it can be helpful, on occasion, not to be burdened by existing knowledge.

Although there may contexts where some *unknown knowns* exist – things you did not realise you knew about a subject – the most important matter to address is the *known unknowns* and the even more elusive *unknown unknowns*. Analytical tactics are needed to help plug these gaps as far, as deep and as wide as possible. You need the capacity to convert *unknowns* into *knowns*. In doing so it will optimise the viewpoint of your subject and better support your judgement about whether to continue with, to refine or to redefine your origin curiosity.

Exploratory Data Analysis

In visualisation, the task of addressing the *unknowns* you may have, as well as substantiating the *knowns* that already exist, involves the use of exploratory data analysis (EDA). This integrates statistical methods with visual analysis to offer ways of extracting wider understanding about what qualities are hidden in your data. We need statistical analysis to describe what is in our data; we need visual analysis to show us what is in our data and, crucially, show us what is not there.

The chart shown in Figure 4.7 is a useful illustration of the value of supplementing stats with visuals. This analysis shows a histogram distribution of the finish times of 9 million marathon

> 'When the data has been explored sufficiently, it is time to sit down and reflect – what were the most interesting insights? What surprised me? What were the recurring themes and facts throughout all views on the data? In the end, what do we find most important and most interesting? These are the things that will govern which angles and perspectives we want to emphasise in the subsequent project phases.'
> **Moritz Stefaner, Truth & Beauty Operator**

runs. The features of this chart follow the classic bell-shape curve that is found in plots about many natural phenomena, such as the height measurements of a large group of people. If we described this data using just statistical analysis we would have found common observations such as the average, maximum, minimums and variance.

However, when you visualise the data, and closely scrutinise some of the patterns, you discover interesting gaps emerging just after the key target finish times on or just before the three-, four- and five-hour marks. You see the shape peaks just before these thresholds and then noticeably drops. This reveals the significance of runners setting themselves and practising for target finish times, often rounded to hourly or half-hourly milestones. This is a quality found in this data that is only revealed by visualising it and only observed through the shape of the results.

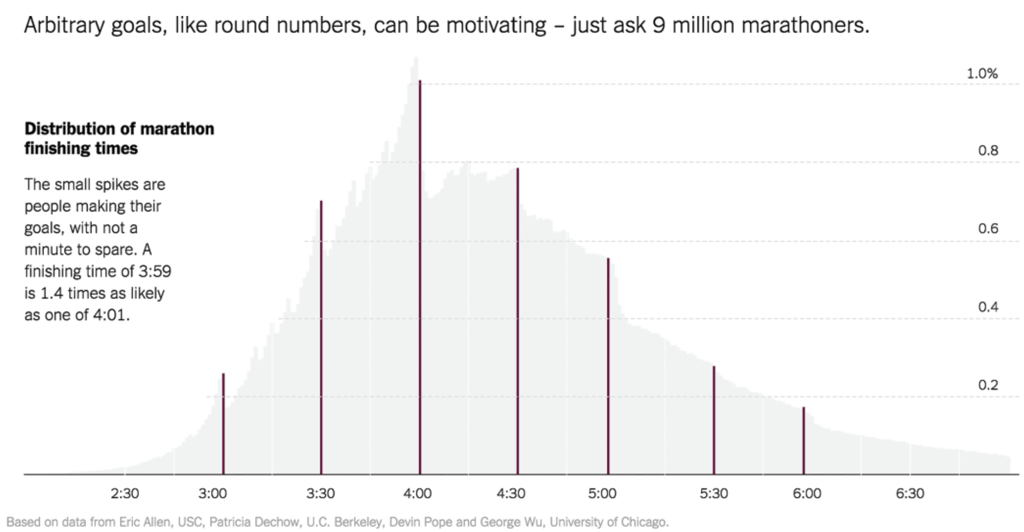


Figure 4.7 What Good Marathons and Bad Investments Have in Common, by Justin Wolfers (*New York Times*)

This example demonstrates the value of EDA, but it does not disclose the secret of how you find such discoveries. There is no instruction manual for this. As John Tukey, the father of EDA, described: 'Exploratory data analysis is an attitude, a flexibility, and a reliance on display, not a bundle of techniques' (1980: 23). There is no single path to undertaking this activity effectively or efficiently; it requires a repertoire of technical, practical and conceptual capabilities, as follows:

Instinct of the analyst: This is the primary matter. The attitude and flexibility that Tukey describes are about recognising the importance of the analyst's traits. Effective EDA is not about the tool. There are many vendors out there pitching their applications as the magic 'point and click' solution to make deep discoveries, and technology inevitably plays a key role in facilitating this endeavour. However, the value of a good analyst cannot be underestimated. It is arguably more influential than the differentiating characteristics between one tool and the next. In the absence of a defined procedure for conducting EDA, an analyst needs to possess the capacity to recognise and pursue the scent of enquiry. A good analyst will have that special blend of natural inquisitiveness and the sense to know what approaches (statistical or visual) to employ and when. Furthermore, when these traits combine with a strong subject knowledge, clearer judgements can be made in distinguishing the significant from the insignificant.

Reasoning: Efficiency is a particularly important aspect of this exploration stage. The act of interrogating data, waiting for it to volunteer its secrets, can take a lot of time and energy. Even with smaller datasets you can find yourself tempted into trying out myriad combinations of analyses, driven by the desire to find an as-yet-undiscovered nugget of golden insight.

There are so many statistical methods and, as you will see, so many visual means for seeing views of data that you simply cannot expect to have the capacity to unleash the full exploratory artillery. In most project circumstances you cannot afford to spend time trying out everything. EDA is about being smart, recognising that you need to be discerning with your tactics.

> 'At the beginning, there's a process of "interviewing" the data – first evaluating their source and means of collection/aggregation/computation, and then trying to get a sense of what they say – and how well they say it via quick sketches in Excel with pivot tables and charts. Do the data, in various slices, say anything interesting? If I'm coming into this with certain assumptions, do the data confirm them, or refute them?' **Alyson Hurt, News Graphics Editor, NPR**

Reasoning can help reduce the size of this prospect. In academia there are two distinctions in approaches to reasoning – deductive and inductive – that I feel are usefully applied in this discussion. Ideally, you will accommodate both approaches to help you confirm your *knowns* and address those elusive *unknowns*:

- *Deductive* reasoning is targeted: You follow a specific curiosity or hypothesis, framed by subject knowledge, and interrogate the data in order to determine whether there is any evidence of relevance or interest in the concluding finding.
- *Inductive* reasoning is much more open in nature: You 'play around' with the data, based on your sense or instinct about what might be of interest, and wait and see what emerges. In some ways this is like prospecting, hoping for that moment of serendipity when you unearth gold. It is important to give yourself room to embark on these somewhat less structured exploratory journeys.

An analogy I often think is useful to help describe EDA concerns a 'Where's Wally?' visual puzzle. The process of finding Wally is unscientific. You often start out by unleashing your eyes

around the scene quite randomly. After this initial burst, perhaps subconsciously, you then adopt a more considered process of visual analysis, eliminating different parts of the scene as 'Wally-free' zones. This aids your focus and informs your strategy for where to look next. As you then move across each mini-scene your eyes are pattern matching, looking out for the giveaway characteristics of the boy wearing glasses, a red-and-white-striped hat and jumper, and blue trousers.

The objective of this task is clear and singular in definition: you know what you are looking for. Unfortunately, the challenge of EDA is rarely that clean, even if you have a source curiosity to shape your pursuit, and you manage to find evidence of your 'Wally' somewhere in the data. In EDA, unlike the 'Where's Wally?' challenge, you have scope also to find other features in your data that might change the definition of what qualifies as an interesting insight. In unearthing other discoveries you might determine that you no longer care about Wally – finding him no longer represents the main enquiry.

> 'I kick it over into a rough picture as soon as possible. When I can see something then I am able to ask better questions of it – then the what-about-this iterations begin. I try to look at the same data in as many different dimensions as possible. For example, if I have a spreadsheet of bird sighting locations and times, first I like to see where they happen, previewing it in some mapping software. I'll also look for patterns in the timing of the phenomenon, usually using a pivot table in a spreadsheet. The real magic happens when a pattern reveals itself only when seen in both dimensions at the same time.' **John Nelson, Cartographer, on the value of visually exploring his data**

> 'My main advice is not to be disheartened. Sometimes the data don't show what you thought they would, or they aren't available in a usable or comparable form. But [in my world] sometimes that research still turns up threads a reporter could pursue and turn into a really interesting story – there just might not be a viz in it. Or maybe there's no story at all. And that's all okay. At minimum, you've still hopefully learned something new in the process about a topic, or a data source (person or database), or a "gotcha" in a particular dataset – lessons that can be applied to another project down the line.' **Alyson Hurt, News Graphics Editor, NPR**

Research: It is important to learn as much as possible about the domain and the data you are working with. Interpreting – the second phase of understanding where you establish meaning – is only possible with domain knowledge. Without this, or having access to resources to help, you will not know if what you are seeing is meaningful. Often, the consequence of EDA is that you simply become more acquainted with the questions you need to ask, rather than any answers.

How to go about addressing this is really just common sense. You need to explore the places (books, websites) and consult the people (experts, colleagues) to give you the best chance of getting accurate answers to the questions you have. Good communication skills are vital. It is not just about talking to others, it is about listening and learning. If you are in a dialogue with subject-matter experts you will have to find an approach that allows you to understand potentially complicated matters and also cut through to the most salient matters of interest.

Nothings: What if you have found nothing? You might reach a dead end, discovering no

significant relationships and finding nothing interesting about the shape or distribution of your data. What do you do then? In these situations you need to change your mindset: *nothing* usually still means *something*. Reaching a dead end or going down blind alleys can be helpful because they help you eliminate dimensions of possible analysis. Traits of nothingness in your data or analysis – gaps, nulls, zeros and no insights – can prove to be the main insight, as described earlier.

There is *always* something interesting in your data. If a value has not changed over time, maybe it was supposed to? That is an insight. If everything is the same size, maybe that is the story? If there is no significance in the quantities, categories or spatial relationships, make the absence of significance your main insight.

Not always needed: It is important to close this discussion about exploration with some pragmatic reality. Not *all* visualisation challenges will involve *much* EDA and not all visualisation projects will give you *space* to do much EDA. Your data might be immediately understandable, and you might have a sufficiently strong knowledge of your subject (lots of *known knowns* already in place). If you are working with small datasets they might not warrant broad visual investigation. You need to be ready and equipped with the capacity to undertake this type of exploration activity when it is needed, but the key point is to judge when it is needed.

Summary: Working with Data

The Four Steps

This chapter commenced your practical involvement with your data, taking you through the four distinct steps that comprehensively acquaint you with the potential of your critical raw material.

Data acquisition looked at the different origins of and methods for accessing your data, including data that is supplied to you, accessed via system download or through web scraping, obtained using an API, gathered through foraging, or involves methods of primary collection.

Data examination profiled the different characteristics that define the type, size and condition of your data. To usefully distinguish different types of data, the 'TNOIR' mnemonic was proposed:

- Textual (e.g. responses to 'Any other comments?' in a questionnaire).
- Nominal (e.g. the gender of a survey participant).
- Ordinal (e.g. the rank of a police officer).
- Interval (e.g. the forecasted temperature in °C).
- Ratio (e.g. the number of sales of a book on Amazon).

Data transformation built on your examination work, identifying ways of modifying and enhancing your data to prepare it for use, including:

- Cleaning: resolve any data condition issues.
- Creating: consider developing new calculations and value conversions.
- Consolidating: think about introducing further data to expand or append to what you already have.

Data exploration discussed the value of using visualisation techniques to supplement statistical approaches as a way to discover more about the qualities and insights hidden away in your data.

General Tips and Tactics

- Perfect data (complete, accurate, up to date, truly representative) is an almost impossible standard to reach, especially given the typical presence of time constraints. You will often need to make a call about when good enough is good enough. You might recognise a point after which your ongoing efforts to refine may result in diminishing returns.
- Do not underestimate the demands on your time; working with data will always be consuming of your attention and effort. Ensure you have built plenty of time into your handling of this data stage; be disciplined by keeping your focus and not getting sidetracked into exploring every possible interesting avenue; be patient and persevere.
- Clerical tasks like file management are important: maintain backups of each major iteration of data, employ good file organisation of your data and other assets, and maintain logical naming conventions.
- Data management practices around data security and privacy will become important when you are working with data that involves more sensitive/confidential subjects.
- Keep notes about where you have sourced data, what you have done with it, any assumptions or counting rules you have applied, ideas you might have for transforming or consolidating, issues/problems, things you do not understand.
- Anticipate and have contingency plans for the worst-case scenarios for data, such as the scarcity of data availability, null values, odd distributions, erroneous values, long values, bad formatting, data loss.
- Ask questions. If you do not know something about your data, do not assume or stay ignorant. And then listen: always pay attention to key information offered by subject-matter experts.
- Attention to detail is of paramount importance at this stage, so get into good habits early and do not cut corners.
- Maintain an open mind and do not get frustrated. You can only work with what you have. If it is not showing what you expected or hoped for, you cannot force it to say something that is simply not there.
- The visuals produced during your data exploration work do not need to be elegantly designed. Do not waste time making your analysis 'pretty', it only needs to inform *you*.

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

5

Establishing Your Editorial Thinking

In Chapter 3, you initiated the design process by expressing the curiosity your work will attempt to satisfy – for you, a stakeholder or your audience. In Chapter 4, you sequentially built up a more intimate understanding of your subject through its data. This may have helped to confirm, revised or expanded the scope of your curiosity.

In the present chapter we move forward to what is, arguably, the least technical stage of the process: establishing your editorial thinking. In the context of a visualisation project, editorial thinking is about determining *what* analysis you are going to portray visually to your audience. The matter of *how* follows next: this stage is the critical bridge between your curiosity definition, your data work and the design steps that follow.

In this chapter you will learn about the importance of editorial thinking in visualisation, what it involves, and how it influences so much of the design thinking that will follow after this stage.

5.1 What is Editorial Thinking?

Editorial thinking is concerned with making informed judgements about the content you intend to include in your visualisation. In my view this is one of the most defining activities that separates the best visualisers from the rest, possibly even more so than any technical talent or design flair. Before we move on to making design choices, you need to consider: given all the things you *could* show, what *will* you show?

There are two words often used to define the essence of editorial thinking. One is editing; the other is opinion. In the context of a data visualisation project, these are both relevant.

Editing is about making selections: choosing what clips you leave in a movie, what contents you leave in a book, how you arrange music into a coherent whole. In visualisation we need to make selections about what analysis we are going to portray to our audience in order to satisfy our articulated curiosity. Regardless of whether origin curiosity still represents our core focus, or if it has evolved following the 'working with data' stage, we will need to decide what analysis will contribute towards facilitating the most relevant understanding about this subject. You can rarely show everything – you rarely should show everything – so what *are* you going to show?

The term *opinion* can imply being impulsive or irrational but, in this context, it means *you* making discerning, informed judgements. The emphasis is clear. It is you who are ultimately responsible for the editing process. Even if it is influenced by guidance you have sought from your project stakeholders or through dialogue with representatives of your audience, every visualisation you ever produced is the consequence of your subjective choices. There is no rulebook for this, no set procedure to lean on, no notion of perfect – it is down to *you* to make the most reasonable call.

You need to decide what you are going to do now because you are about to move towards the design phase that will involve picking chart types, deciding on which colours to use, what layout to construct, and more besides. Before you decide how to design your content, you have to determine what content to include.

> 'A photo is never an objective reflection, but always an interpretation of reality. I see data visualisation as sort of a new photojournalism – a highly editorial activity.' **Moritz Stefaner, Truth & Beauty Operator**

When explaining what it means to establish editorial thinking in practice I find it helpful to consider the parallels that exist between data visualisation and photography (or perhaps more specifically, photojournalism, when there is a more explicit aim to communicate and report). Think of a chart as a photograph of your data. By considering some of the decisions involved in taking a photograph, you will find useful perspectives to shape your editorial thinking in visualisation. There are three particular perspectives to consider: angle, framing and focus.

Angle: The first aspect of editorial thinking is concerned with choosing the angle of analysis – the view of your data – that you think will best support the understanding required for this subject. What content will answer or contribute towards answering the overriding curiosity?

In photography the angle is the position from which you take a shot and the view of your subject this position gives you. Just as with any photograph, a visualisation is limited in that it cannot provide a panoramic 360° view of all your data simultaneously. There is only so much a single chart can show.

So, what are you going to show? Will it be how values have changed over time, or how they look spatially, on a map? Is it more important to show a categorical breakdown or portray important relationships between different variables?

Each different chart you will meet in the next chapter provides a different view of a subject through its portrayal of your data. Before you choose to use a chart, you need to nail down what angle of analysis you want to provide. Furthermore, will one angle be sufficient, or might you need several different views?

It is easy to find yourself reluctant to commit to just a singular choice of angle. Even in a small dataset, there are typically multiple possible angles of analysis you could conduct. It is often hard to ignore the temptation of wanting to include multiple angles to serve more people's interests. I often find it far too easy to see everything as being potentially interesting to my audience, the curse of the analyst.

It is important, though, not to fall into the trap of lazy thinking that if you throw together multiple angles of analysis into your work, eventually one will serve the interests of your audience. Just because you take 100 photographs of your holiday does not mean you should show them all to somebody. You need to demonstrate discipline.

Let's suppose you are becoming increasingly active as a runner and are curious about how well you are doing, drawing from run data you have been collecting using a tracking device or smart watch which you can download and analyse yourself.

> 'I think this is something I've learned from experience rather than advice that was passed on. Less can often be more. In other words, don't get carried away and try to tell the reader everything there is to know on a subject. Know what it is that you want to show the reader and don't stray from that. I often find myself asking others "do we need to show this?" or "is this really necessary?" Let's take it out.' **Simon Scarr, Deputy Head of Graphics, ThomsonReuters**

Your core curiosity may be expressed as 'How well am I running?' This is a rather open-ended enquiry and is not going to be answerable by a single statistic or one single view of data. Even if, after a specific run, there is only one thing dominating your attention at that specific moment (such as 'what distance did I go today?'), after each subsequent run there might be different perspectives of interest as your focus shifts. To answer this ongoing curiosity, you will need access to several distinct angles of analysis that, when synthesised, will collectively provide the understanding you need, as listed in Figure 5.1.

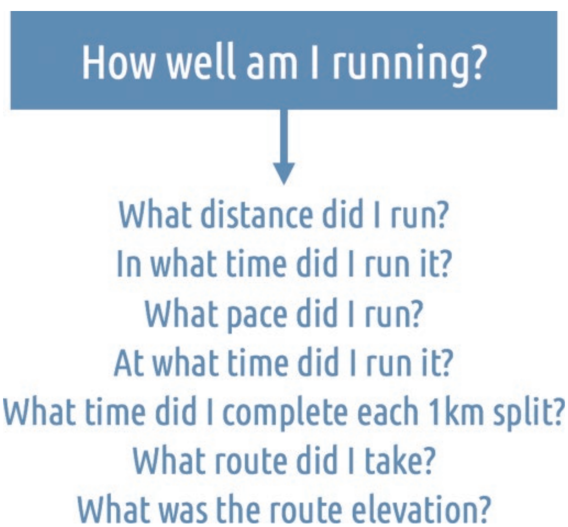


Figure 5.1 The Questions That Might Help Answer the Query, 'How Well Am I Running?'

As explained in the first chapter when discussing articulating your curiosity, I find forming data questions helpful at this point. It keeps me focused on what I am trying to answer, albeit now at a degree of increased specificity. All the questions listed in Figure 5.1 reflect reasonable contributors towards collectively answering the curiosity. Some of these questions will be answered by individual statistics, some will require charts to show an answer.

Framing: After defining which angle or angles of analysis you might need to include, framing is the second editorial perspective concerned with refining the contents to be included in your analysis. In photographic parlance this relates to choices about the field of view: what will be included inside the frame of the photograph and what will be left out? Just like a photographer, a visualiser must demonstrate careful judgement about what data items to include, what data items to exclude, and why, for each statistical or chart display. A balance must be struck to find the most representative view of your content.

They say the camera never lies, but photographs can certainly be distorting. With visualisations, if you filter off too much of the content, it might disguise important context required for interpreting the significance and meaning of values. Conversely, if you avoid filtering your content you may fail to make visible the most salient discoveries. One of the key motives of framing is to remove unnecessary clutter – there is only so much that can be accommodated in a single view and only so much your audience will be able to process. Do not give them a puzzle if you can give them the answer.

The criteria for your framing judgements will be influenced by the amount of data available to show and the complexity of what you want to portray. Sometimes, showing lots of things – indeed, *all* the things – creates visual complexity and that actually supports the point you are making ('look how crazily complex this system is!'). Further considerations about the setting, such as the need for rapid insights or scope for deeper, more prolonged engagements, and the dimensions and medium of the output, will also have a bearing on this matter.

Returning to the scenario of running data, the editorial framing decisions about this may cover the following:

- Do I include only my latest run, other recent runs or all my runs?
- Do I include only those runs where I covered a certain minimum distance?
- Do I include only runs where I have used specific routes or all routes?
- Do I include only my data in this analysis or are there other people whose data I can compare (e.g. a partner or running mate)?

Focus: This is the third editorial perspective and concerns which items of your data you might choose to emphasise, thus generating some level of focus for the attention of your viewer. This decision is not a function of filtering – that is the concern of your framing decisions. Focus is about subjectively choosing to contrast visually features of your display that you deem to be more important than others.

The best photographs are able to balance light and colour, not just to set the mood of a subject, but to help illuminate key elements and convey visual depth. In addition to the depiction of the relative sizing and arrangement of different contents, this provides a sense of visual hierarchy to direct the eye of the viewer.

Whereas framing judgements were about balancing the clutter of your content, this is about balancing the volume. If everything is shouting, nothing is heard; if everything is in the foreground, nothing stands out; if everything is bold, nothing dominates.

The relevance of editorial focus is primarily associated with explanatory visualisations, whereby elevating key insights to the surface of a display is a key attribute of the experience they provide. So, what features of your data need to be brought into the foreground, left in the mid-ground, or relegated into the background? What needs to be bigger and more prominent and what needs to be less so?

For your running data, you might use colour highlighting to emphasise your above-average runs and add labels to point out the fastest and the furthest. It will depend on what features you are focusing on.

5.2 The Influence of Editorial Thinking

It is useful to ground this discussion practically by explaining how these editorial perspectives will affect your design thinking. This is not just going to influence the best choice of chart to represent your angle of analysis, but may influence the way you employ interactivity, the annotations you include, the colours you apply, and the composition of your content. Let's look at two examples that illustrate this connection of influence between editorial and design thinking.

'The Fall and Rise of US Inequality'

The first example (Figure 5.2) is a chart taken from an article published on the 'Planet Money: The Economy Explained' section of the US-based National Public Radio (NPR) website. The article is titled 'The Fall and Rise of US Inequality in 2 Graphs'. As the title suggests, the full article includes two charts, but I just want to focus on the second of these for the purpose of this illustration.

Let's assess the editorial perspectives of angle, framing and focus as demonstrated by this work.

Angle: The main angle of analysis can be expressed as: 'What is the relationship between two quantitative measures (average income for the bottom 90% and for the top 1% of earners) and how has this changed over time (year)?' This angle would be considered relevant because the relationship between the *haves* and the *have-nots* is a key indicator of wealth distribution. It is a topical and suitable choice of analysis to include with any discussion about inequality in the USA.

Framing: The parameters that define the inclusion and exclusion of data in the displayed analysis involve filters for time period (1917 to 2012) and country (just for the USA). The starting point of the data commencing from 1917 may just be an arbitrary cut-off point or could be a more significant milestone in the narrative. More likely, it probably represents the earliest available data. The up-to-date-ness of any chart can expire immediately after publication, but although this data only reaches as far forward in time as 2012 (despite publication in 2015), the analysis is of such historical depth that it should be considered suitably representative of the subject matter. To focus only on the USA is entirely understandable given the scope of the piece.

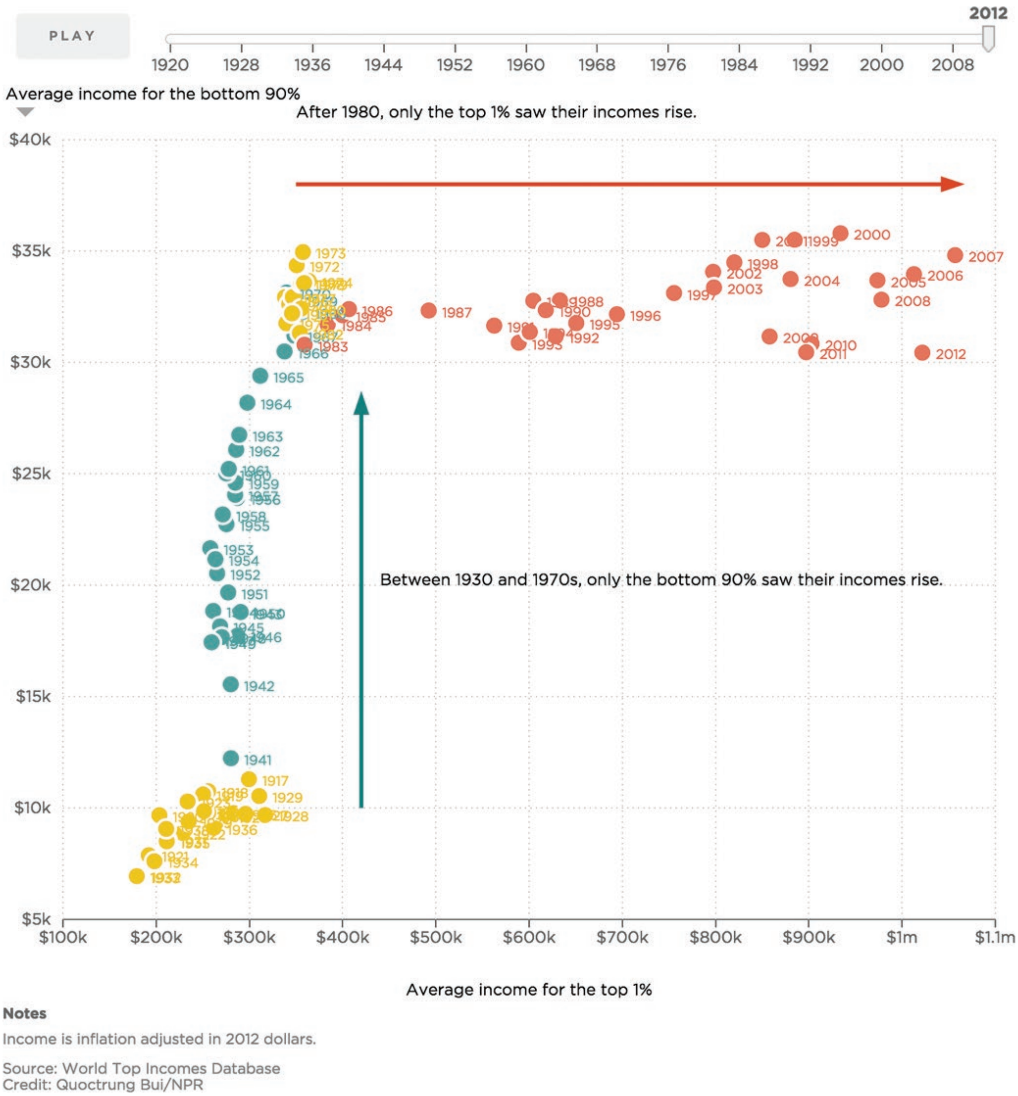


Figure 5.2 The Fall and Rise of US Inequality, in 2 Graphs

Source: World Top Incomes Database; Design credit: Quoctrung Bui (NPR)

Focus: The visualisation includes an interactive 'time slider' control that allows users to move the focus incrementally through each year, colouring each consecutive yearly marker for emphasis. The colours are organised into three classifications to draw particular attention to two main periods of noticeably different relationships between the two quantitative measures.

Now let's switch our view and have a look at how the five layers of design are influenced by this editorial thinking.

Data representation: The *angle* is what fundamentally shapes the data representation approach. In simple terms, it determines which chart type is used. In this example, the desired angle of analysis is to view the relationship between two quantitative measures over time (average

income for bottom 90% vs top 1% of earners). A suitable chart type to portray this visually is the *scatter plot*, as selected. You will learn in the next chapter that the scatter plot belongs to the 'relational' family of chart types in that it primarily displays the relationships that might exist or otherwise between different variables. Given there was also a dimension of time expressed in the intended angle of analysis, a chart type from the 'temporal' family of charts *could* have been used. However, with the main emphasis being to show the relationships, the scatter plot was the better choice. The *framing* perspective defines what data will be included in the chosen chart: as mentioned, only data for the USA and the time period 1917–2012 is displayed.

Interactivity: As you will discover in Chapter 7, the role of interactivity is to enable adjustments for *what* data is displayed and *how* it is displayed. The sole feature of interactivity in this project is offered through the 'time slider' control, as mentioned, which sequences the unveiling of the data points year by year in either a manual or automated fashion. The inclusion of such interactivity can be influenced by the editorial *focus*, in this case to unveil the yearly values in sequence emphasising the position of each consecutive value over time.

Annotation: The primary chart annotations used here are the two arrows and associated captions, drawing attention to the two prominent patterns that characterise a fall and a rise in inequality. The inclusion of these captions would be a consequence of editorial *focus* to determine that these patterns in the data should be surfaced for the viewer.

Colour: As you will learn in Chapter 9, one of the key applications of colour is to create ordinal emphasis that brings important content to the surface and draws the eye's attention. This would influence the decision to deploy four colour states within the chart: a neutral colour to show all points at the start of the animation and then three different emerging colours used to separate the three clustered groups visually. The absolute colour choices of red, green and yellow tones are not directly informed by editorial thinking, rather it is the identified need to have four different colours to draw out the key patterns.

Composition: This element of design concerns decisions about all of the positioning and sizing decisions. In this example, editorial thinking will have had limited influence over the composition choices in this chart.

'Why Peyton Manning's Record Will Be Hard to Beat'

In this second example, published in 'TheUpshot' section of the *New York Times* website, there are three charts presented in an article titled 'Why Peyton Manning's Record Will Be Hard to Beat'. Here I will look at all three charts.

Let's once again start by assessing the editorial perspectives of angle, framing and focus as demonstrated by this work, one chart at a time.

Angle: The first chart (Figure 5.3) portrays an angle of analysis that can be expressed as: 'How have quantitative values (cumulative NFL touchdown passes) changed over time (year) for multiple categories (quarterbacks)?' This analysis was relevant at the time due to the significance of Peyton Manning setting a new record for NFL quarterback touchdown passes, an historic moment, and, according to the article, 'evidence of how much the passing game has advanced through the history of the game'. Inspired by this achievement, the question posed

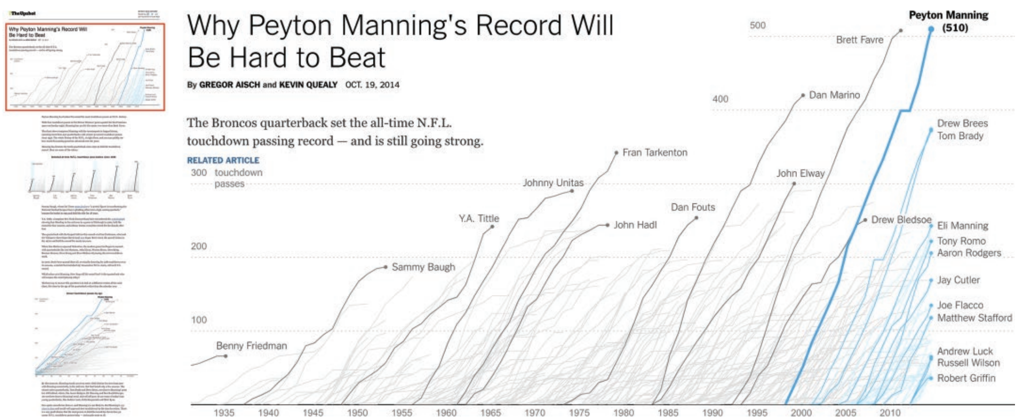


Figure 5.3 Why Peyton Manning's Record Will Be Hard to Beat, by Gregor Aisch and Kevin Quealy (*New York Times*)

by this article overall is whether the record will ever be bettered – which would have likely been the origin curiosity that drove the visualisation project in the first place. On its own, this analysis would be deemed insufficient to support the overarching enquiry, as evidenced by the inclusion of two further charts.

Framing: The criteria for data inclusion are shaped by the time period (1930 to 19 October 2014) and qualifying quantitative threshold (minimum of 30 touchdown passes). They are representative of the truth at the moment of publication (i.e. up to 19 October 2014), though clearly the data would no longer be up to date as soon as the next round of games took place. At the time of publishing this book, Manning's record is likely to be surpassed by either Tom Brady or Drew Brees, and possibly both. This reinforces the idea of a chart being a frame in time. The inclusion of players with at least 30 touchdown passes would be informed either by knowledge of the sport (and if 30 touchdowns were seen as a common threshold) or possibly from discovering that this was a logical cut-off value having visually explored the shape of the data for *every* quarterback.

Focus: There is editorial emphasis applied to highlight the record holder as well as distinguishing a selection of other current players. This helps to see which other contemporary players (at the time) *could* have a chance of pursuing this record. Knowing what we know now, it was not unreasonable to expect Brady and Brees to be among the main candidates to pursue Manning's record. There is also further emphasis applied to contrast selected all-time NFL touchdown pass leaders with every other qualifying player.

Angle/Framing: In the second chart (Figure 5.4), the same editorial perspectives apply for *angle* and *framing*, but the *focus* has changed. Even though the charts now comprise several small, repeated views, each one focusing on the career trajectories of a selected previous record holder, it is still fundamentally the same underlying view of data and includes the same data.

Focus: Colour is used to emphasise the previous record-holding players' career line in each chart panel, with background colour banding used to illuminate the duration of their record standing. Value labels reveal the number of touchdowns achieved by each.

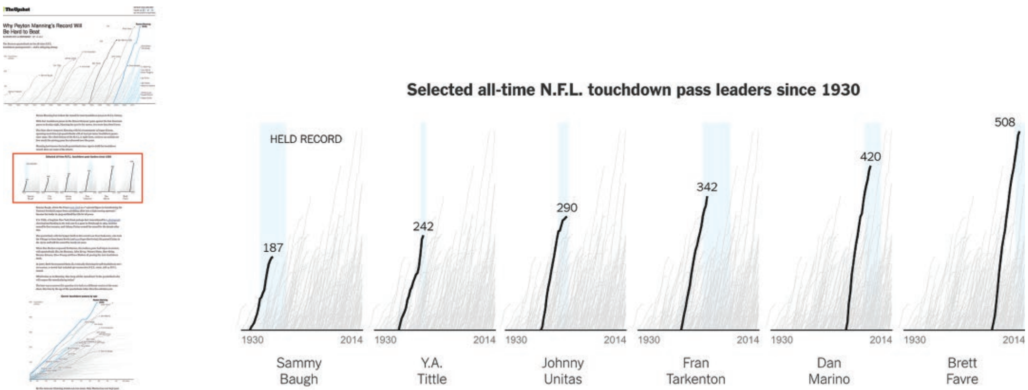


Figure 5.4 Why Peyton Manning's Record Will Be Hard to Beat, by Gregor Aisch and Kevin Quealy (*New York Times*)

The third and final chart (Figure 5.5) has many similarities with the first (Figure 5.3). Once again it maintains the same consistent definition for *framing* and it has the same *focus* as the first chart, but now there is a subtle difference in *angle*.

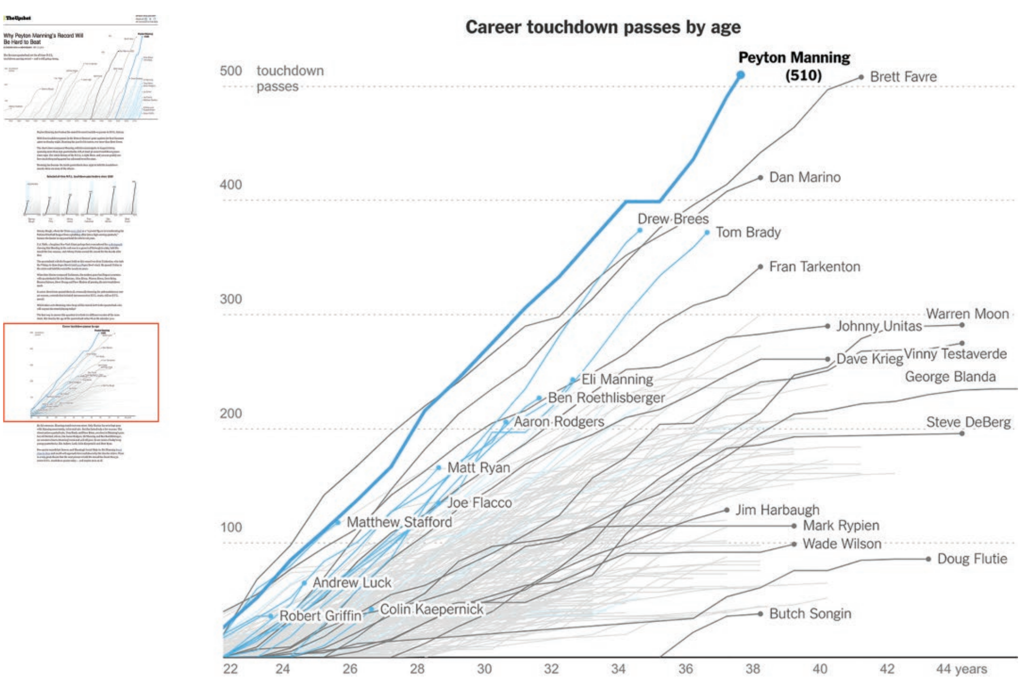


Figure 5.5 Why Peyton Manning's Record Will Be Hard to Beat, by Gregor Aisch and Kevin Quealy (*New York Times*)

Angle: The third and final chart (Figure 5.5) has many similarities with the first (Figure 5.3), but the view of data now shown would be expressed as: 'How have quantitative values (cumulative NFL touchdown passes) changed over time (age) for multiple categories (quarterbacks)?' The difference is the temporal measure plotted along the x-axis and is now about the players' ages at the time of each touchdown pass, rather than when it occurred. This is a small but relevant difference as it changes the nature of the analysis. It is included in support of the enquiry posed about whether 'the quarterback who will surpass Manning's record is playing today?' Incidentally, the article concludes it is going to be a very difficult record to beat.

Framing/Focus: This chart maintains the same defined perspectives for *framing* and *focus* as the first chart.

Let's now switch our view and have a look at how the five layers of design are influenced by this editorial thinking.

Data representation: As I have stated, the *angle* and *framing* dimensions are hugely influential in the reasoning of chart-type requirements. In each of the charts used we are shown different perspectives around the central theme of how touchdown passes have changed over time for each qualifying quarterback. A line chart was an entirely appropriate way to show the trends of cumulative touchdown values for all the players included. Not surprisingly, the line chart belongs to the 'temporal' family of chart types, as you will see in Chapter 6. Alternative angles of analysis may have been possible to pursue, such as exploring the relationship between the age of players when they reached their highest total and the absolute total touchdown passes. For this analysis, a scatterplot would have been ideal to show this. However, showing the stories of each player's trajectory towards their cumulative touchdown total made for a more compelling display.

Interactivity: The only feature of interaction is included in the first and third charts, offering mouseover events to reveal annotations of the names and total passes of any of the players presented as grey lines. This serves the appetite of any viewer curious about the names of those players without a label, but also preserves a certain elegance by not over-cluttering the main chart display: detail is only available on demand, one player at a time.

Annotation: The interactive-enabled labelling is effectively a joint matter concerned also with annotation. The decision to include permanent annotated value labels in each chart provides editorial emphasis (in the first and third charts) for Peyton Manning, selected current quarterbacks and selected all-time NFL touchdown pass leaders. The second chart only provides single-value labels in each chart panel for the respective record holder on display.

Colour: The approach to creating focus is further amplified using colour. In the main chart, emphasis is again drawn to Peyton Manning's line, as the record holder (thick blue line), other current players (highlighted with a blue line) as well as selected all-time NFL touchdown pass leaders (dark-grey line). For the second chart the light-blue-coloured banding draws out the period of the records held by selected players down the years. This helps the viewer to perceive the duration of their records.

Composition: The sequencing of the charts in the article is a function of editorial *focus* – what should go first, second and last, and why? Given the limitations of screen space to consume this article, the ordering of the charts in this way will be to support the main narrative and essentially answer the curiosity of the piece, as expressed in the title.

Determining Relevance

So, you can see how editorial definitions influence the design thinking that follows. But how do you arrive at these definitions? What determines if you have astutely identified the right editorial perspectives?

In Chapter 2, we discovered that one of the key principles for good visualisation design is that it should be accessible. A characteristic of fulfilling accessibility in design was relevance, specifically a concern about whether you are providing your audience with access to the most useful understanding about this subject.

A lack of relevance is a curse that strikes a lot of visualisation work. Turning data into a visual just because you happen to have it available is an aimless exercise. That is why we need to instigate from the origin of a curiosity. This should provide a reasonably informed view about what could be most useful to your audience, at least initially. Now that you have spent more time deliberating over your audience's needs, maybe even asking them what they need to know, it is time to determine what is truly useful to them. This involves a blend of considerations:

- Timeliness: Is the understanding beneficial at the moment of encounter?
- Interestingness: Is the topic stimulated by new understanding ('man bites dog!') or representative of helping to reinforce existing understanding ('dog bites man!')?
- Pertinence: Does the viewer have an established association with the topic?
- Sufficiency: Is the level of detail provided appropriate to the viewer's needs at the moment of encounter?

Usefulness as perceived by the audience is not the only factor that shapes relevance, though. It is also informed by what you actually have available to show them (Figure 5.6): what data do you have and what analysis is available for you to present? If you do not have *it* in your data, you cannot show *it*.

> 'It requires the discipline to do your homework, the ability to quiet down your brain and be honest about what is interesting.' **Sarah Slobin, Visual Journalist**

Having gone through the process of examining and exploring your data, in particular, you will now have a far more informed view about what you could show your audience in order to meet all their needs of usefulness. Indeed, sometimes your audience will not really be best placed to know what is useful to them, in which case you may lead on what *you* want your audience to know. Depending on the context, and your proximity to the subject and its data,

you might have the autonomy to dictate what you want to say, more so than what you think the audience want to see.

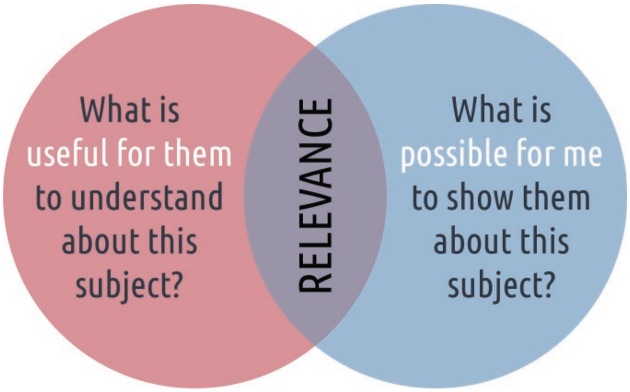


Figure 5.6 An Illustration for Determining 'Relevance'

It is vital not to fall into the trap of going through the motions. Just because you have spatial data does not mean that the most useful angle of analysis will concern the 'where' of your data. If the interesting insights about that subject are not significantly influenced by the spatial dimension, a map may not provide the most relevant window on that data. You might actually find the location information is more useful as a categorical device to group or separate analysis and so other forms of analysis may be more insightful to pursue.

Although presented as consecutive stages, 'working with data' and 'editorial thinking' are quite iterative: working with data influences your editorial perspectives; and your editorial perspectives in turn may influence activities around working with data. In practice there will be much toing and froing between the two (in contrast to the linear way I have to write and present this book). The data transformation activity, in particular, is a key link. Editorial definitions may trigger the need for more data to be gathered about the specific subject matter or further consolidation to support the desired angles of analysis or the framing dimensions. Editorial definitions might also influence the need for further calculations, groupings or general modifications to refine the preparedness of your data for displaying the analysis.

Summary: Establishing Your Editorial Thinking

Editorial Perspectives

In this chapter you reflected on the possibilities offered by your data and learned about the importance of committing to an editorial path. You defined three key editorial perspectives that should be relevant to your audience in support of the overriding curiosity you are pursuing:

- Angle: What view(s) of your data is most relevant? In language terms, what question should your eventually chosen charts answer?

- Framing: What data items and values will you include and exclude? What is most representative of your subject?
- Focus: Are there any features of your data you would wish to emphasise? This is especially relevant to explanatory visualisations: if you have something to say, say it.

General Tips and Tactics

- If your data is riddled with data condition issues, such as gaps or errors, perhaps consider making *this* the story: invert the editorial view to be about the data, rather than about the subject through its data.
- There is *always* something interesting in your data: you just might not be equipped with sufficient domain knowledge to know this or it may not be currently relevant. Get to know the difference between relevant and irrelevant by researching and communicating to learn more about your subject.
- To help with your editorial angle, think about what title you would attach to this work. What would be the headline? What would be the question posed that the visualisation might answer?

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

Part C

Developing Your Design Solution



6

Data Representation

In Chapters 3, 4 and 5 you have been working through activities that embody what I consider to be the hidden thinking of a visualisation project. These preparatory stages have helped you define the requirements and aims of your work, given you steps to become acquainted with your data, and, most recently, provided a structure for defining your editorial intent.

This chapter commences the fourth stage of the design process and represents a shift in focus towards design thinking. 'Developing your design solution' begins with arguably the most significant element of the visualisation design anatomy, namely data representation. How will you visually portray your data?

We start the discussion by looking at the fundamentals of visual encoding, exploring the building blocks that underpin all data representation thinking. From this bottom-up viewpoint we will switch to the more pragmatic perspective of selecting chart types. To close the chapter, you will learn about the influencing factors that will inform the choices you make.

6.1 Visual Encoding and Charts

Representing your data visually involves the act of visual encoding. As visualisers, we encode our data using two main visual properties, marks and attributes. *Marks* are visual placeholders representing data *items*, such as distinct records or discrete groupings, depending on the form of your tabulation. These are the four main types of marks, as shown in Figure 6.1.

Attributes are variations in the visual appearance of marks to represent the values associated with each data item. The main attributes you will encounter include those given in Figure 6.2.

The creative scope of some projects may use variation in attributes around the auditory (sound), haptic (touch), gustatory (taste) and olfactory (smell) senses, otherwise these visual attributes are the most commonly used options.

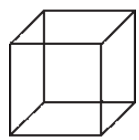
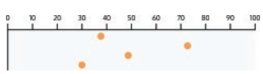
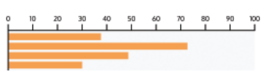
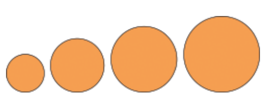
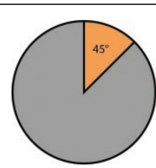
MARK	EXAMPLE	DESCRIPTION
Point		The *point* mark is commonly used as a marker to represent quantitative values through position on a scale, forming the basis of, for example, the scatter plot.
Line		The *line* mark is commonly used to represent quantitative values through variation in size (length), forming the basis of, for example, the bar chart.
Shape		The *shape* mark is commonly used to represent quantitative values through variation in size and position, forming the basis of, for example, the bubble plot.
Form		The *form* mark is used to represent quantitative values through variation in size (volume), forming the basis of charts that encode 3D representations.

Figure 6.1 A Classification of Different Types of 'Mark' Encodings

ATTRIBUTE	EXAMPLE	DESCRIPTION
Position		Variation in position along a scale is used to indicate a quantitative value, often using a point mark.
Size (Length)		Variation in size (length) is used to represent quantitative values based on proportional scales where the larger sizes mean larger quantities. The line mark has a single 'linear' spatial dimension, i.e. it shows quantities through either height or width but not both.
Size (Area)		Variation in size (area) is used to represent quantitative values based on proportional scales where the larger sizes mean larger quantities. The shape mark has two ('quadratic') spatial dimensions i.e. it shows quantities through a combination of both height and width.
Size (Volume)		Variation in size (volume) is used to represent quantitative values based on proportional scales where the larger sizes mean larger quantities. The form mark has three ('cubic') spatial dimensions i.e. it shows quantities through a combination of height, width, and depth.
Angle		Variation in the size of an angle is used to represent quantitative values where larger angles mean larger quantities or, more specifically, larger parts of a whole.

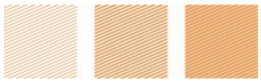
ATTRIBUTE	EXAMPLE	DESCRIPTION
Quantity		Variation in the quantity of a set of point marks (such as symbols) can be used to represent a single or aggregated quantitative value.
Colour: Hue		Variation in colour hue is typically used for distinguishing categorical data values.
Colour: Saturation		Variation in colour saturation can be used, often in conjunction with other colour properties, to represent ordinal scales; typically, the greater the saturation, the greater the hierarchical emphasis.
Colour: Lightness		Variation in the lightness of colour can be used to represent quantitative scales; typically, the darker the colour, the higher the quantity.
Pattern		Variation in pattern (sometimes also described as pattern *texture* or *density*) can be used to represent ordinal scales or distinguish categorical values, perhaps indicating degrees of certainty.
Symbol		Variation in symbols are commonly used for distinguishing categorical data values. The scope of this attribute could extend to images and illustrations explicitly representative of data values.
Connection		Connection (also known as *edge*) indicates a relationship between two nodes established by a connecting line. The shape and size of the connection is usually meaningless but sometimes arrows or variation in line thickness may be used to encode some notion of direction in the relationship.
Containment		Containment (also known as *enclosure*) is a way of encoding a hierarchical relationship between categories that belong to a related 'parent' category grouping.

Figure 6.2 A Classification of Different Types of 'Attribute' Encodings

It is worth noting that sometimes you do not need to encode data. Displaying values in their original numeric or textual form may suffice, perhaps as presented in a table or through callout statistic headlines.

Understanding visual encoding is of fundamental importance and is of particular relevance when representing data using tools that adopt a bottom-up approach. However, for most people's needs, it can often be more pragmatic to think about data representation techniques through selecting chart types.

CATEGORICAL | Comparing categories and distributions of quantitative values

HIERARCHICAL | Revealing part-to-whole relationships and hierarchies

RELATIONAL | Exploring correlations and connections

TEMPORAL | Plotting trends and intervals over time

SPATIAL | Mapping spatial patterns through overlays and distortions

Figure 6.3 The 'CHRTS' Families of Chart Types

If marks and attributes are the ingredients, chart types are the recipes. Different charts offer different established ways of representing data, each one comprising combinations of marks and attributes. As the field has matured, and practitioners have developed new recipes of marks and attributes, the range of established chart-type options has grown.

To acquaint you with a broader repertoire of charting options, over the coming pages I will present a collection of some of the common and useful chart types being used across the field today. This gallery aims to provide you with a valuable reference that will help you to decide how best to show what it is you want to say. I have organised each chart into five main families (Figure 6.3) based on the primary editorial relationship you are trying to understand. The five-letter mnemonic CHRTS provides a useful taxonomy for organising your thinking about which chart(s) to use for your data representation needs.

Each chart-type profile is presented with supporting details that will help you fully understand the role and characteristics of each option, including:

- The primary name used to label each chart type as well as some further alternative names that are often used.
- An indication of which CHRTS family each chart belongs to, based on their specific primary role, as well as a sub-family definition for further classification.
- A description of the chart's representation method, detailing what it shows and what each mark and attribute encoding it deploys.
- An applied example of the chart type in use with a description of what it specifically shows.
- Presentation tips about the potential interactivity, annotation, colour or composition design choices you might consider.
- 'Variations and alternatives' that describe further derivatives to understand other uses and different purposes.

This gallery of charts is by no means an exhaustive list and I have excluded some options because they were not different enough from other charts that have been profiled. I have mentioned some charts that are legitimate derivatives or alternative applications of other similar charts, but have assigned a whole page to profile these separately. For example, the *Voronoi treemap* is really just a variation on the treemap that is profiled. It uses a different

algorithm to arrange its constituent pieces within different spatial layouts, like circles. The appearance and method of making this might be slightly different, its usage is not.

I have wrestled with the value of including some of the charts presented, often due to limitations and shortcomings in aspects of their usage. Some charts have merit for specific contexts, but can be quite narrow in scope. Therefore, by including certain partially flawed charts I am attempting to signpost relevant shortcomings, so you know how to use them sparingly. A word cloud, for example, is a chart with absolutely quite limited value, but nonetheless it does have a role, as does the often-derided pie chart. All chart types offer value for different situations; you just need to use discretion to select them only under specific circumstances.

Although I have excluded several charts on grounds of demonstrating only a slight variation on profiled charts, there are some types included that do exhibit small derivations from other charts (such as the bar chart and the clustered bar, or the scatter plot and the bubble plot). In these cases I felt there was sufficient difference in their practical application, and they were in common usage, to merit their separate inclusion, despite the similarities.

Another point to make is that certain charts do not just fit into a single family. All charts that belong to the hierarchical, relational, temporal and spatial families can include features of categorical breakdown. Using a line chart to show how quantitative values have changed over time for different categories could warrant being classified in either the temporal or the categorical families. However, the change over time dimension is the primary dimension of analysis and enables comparison between categories as a secondary perspective, so it is assigned to the temporal family. I have therefore concentrated the taxonomy around the angle of analysis each chart primarily conveys.

Finally, the spatial family of charts often relates to thematic maps that would not normally be considered charts in purist terms. For convenience, though, I am badging them all as charts. It is worth noting too that not all spatial analysis is geographic. Any of the spatial methods presented could be used for non-geographic contexts, such as the anatomy of the body, the layout of a building, the seat plan of an airliner.

BAR CHART

ALSO KNOWN AS Column chart, histogram (wrongly), lollipop chart

C H R T S

COMPARISONS

REPRESENTATION DESCRIPTION

A bar chart displays quantitative values for different category items. The chart comprises line marks (bars) with the size attribute (length or height) used to represent the quantitative value for each item.

EXAMPLE Comparing the number of unique land neighbours for countries with at least seven.

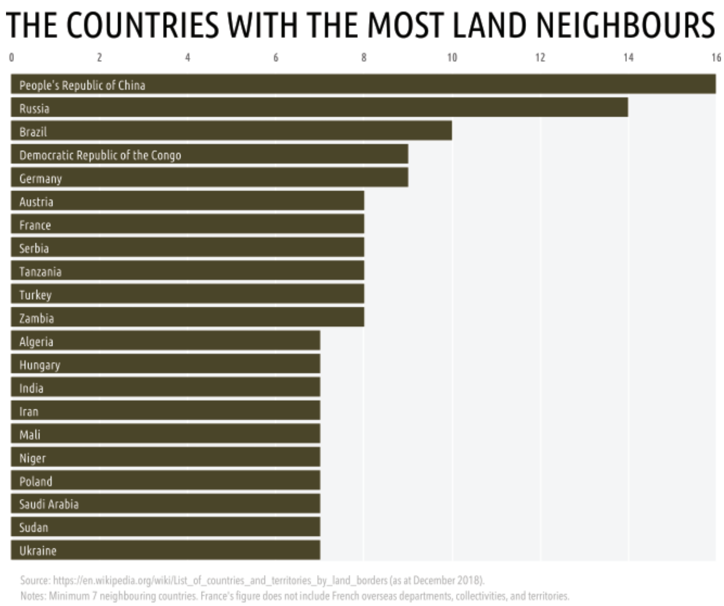


Figure 6.4 The Countries with the Most Land Neighbours

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. If you include axis-scale labels you should not need to label directly each bar value, as this will lead to label overload.

COMPOSITION: The bars should be proportionally sized according to the associated quantitative value – nothing more, nothing less – otherwise the perception of the bar sizes will be distorted. Most commonly, this means setting the quantitative value scale to an origin of zero. There is no significant difference in perception between vertically or horizontally arranged bar charts; it will depend on which layout makes it easier to accommodate the range of values and to read the item labels associated with each bar. Including a small gap between each bar will help to preserve a clear distinction between each category item. Aim to make the sorting of values in the chart as meaningful as possible.

VARIATIONS & ALTERNATIVES

A variation in the application of a bar chart would be to show quantitative values over time. This would be an option to consider over the line chart when you have quantities for discrete periods (such as totals over a monthly period) rather than a purely continuous series of point-in-time measurements. 'Spark bars' are mini bar charts that aim to occupy only a word's length amount of space. They are often seen in dashboards where space is at a premium and there is a desire to optimise the density of the display. If you want to include further categorical subdivisions, an alternative might be the 'clustered bar chart', to compare two or more or adjacent values, or the 'stacked bar chart', if there is a part-to-whole relationship. 'Dot plots' offer a useful alternative for situations where you have large quantitative values with a narrow range of difference and this difference is important to make visible. For contexts where you have diverse value sizes and many categorical items, the 'proportional symbol chart' is an option to consider.

CLUSTERED BAR CHART

C H R T S

COMPARISONS

ALSO KNOWN AS Clustered column chart, paired bar chart, grouped bar chart

REPRESENTATION DESCRIPTION

A clustered bar chart displays quantitative values for different primary category items with a secondary categorical breakdown enabling local comparisons. The chart comprises line marks (bars) with the size attribute (length or height) used to represent the quantitative value for each item. An attribute of colour is also used to distinguish further the secondary categorical groupings.

EXAMPLE Comparing the number of Oscar nominations with the number of Oscar awards won for the 10 actors who have received the most nominations for acting.

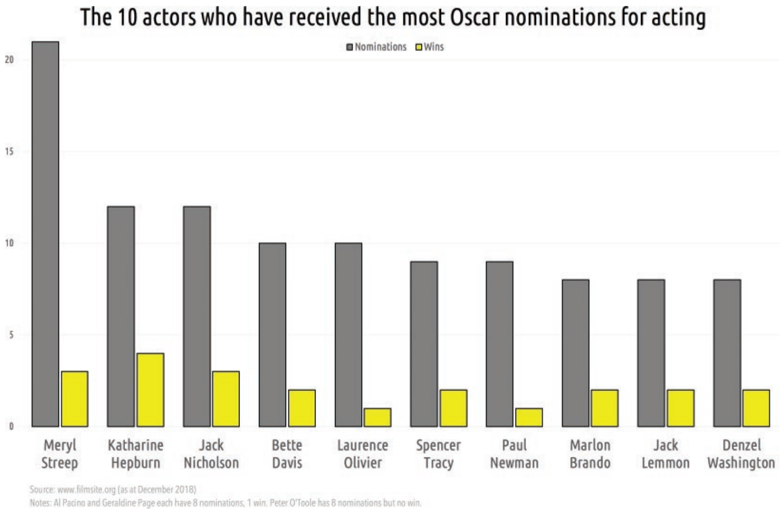


Figure 6.5 The Ten Actors Who Have Received the Most Oscar Nominations for Acting

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. If you include axis-scale labels you should not need to label directly each bar value, as this will lead to label overload. Any colours used must be explained through the inclusion of a legend.

COMPOSITION: The bars should be proportionally sized according to the associated quantitative value – nothing more, nothing less – otherwise the perception of the bar sizes will be distorted. Most commonly, this means setting the quantitative value scale to an origin of zero. There is no significant difference in perception between vertically or horizontally arranged clustered bar charts; it will depend on which layout makes it easier to accommodate the range of values and to read the item labels associated with each cluster. Including a noticeable gap between each cluster of bars will help to preserve a clear distinction between each primary category item. Sometimes one bar might be slightly hidden behind the other if the display concerns a before and after relationship. Aim to make the sorting of values in the chart as meaningful as possible.

VARIATIONS & ALTERNATIVES

Like the bar chart, clustered bar charts can also be used to show how values have changed over time. Alternatives would include the 'connected dot plot', particularly to compare the quantitative size of two categories across a number of major category items. If your clusters comprise many distinct categories, the display might become too busy. You therefore might consider creating separate bar charts for each category item or using a 'matrix chart' structure to show the quantitative values at the intersection of two categorical dimensions.

BULLET CHART

ALSO KNOWN AS {No other names}

REPRESENTATION DESCRIPTION

A bullet chart is effectively a bar chart displaying quantitative values for different categories, but incorporating additional bandings to assist with interpreting the bars. The chart comprises line marks (bars) with the size attribute (length or height) used to represent a quantitative value for each item. An attribute of colour (usually the lightness property) is commonly used to distinguish contextual bandings behind each bar to aid interpretation.

EXAMPLE Comparing the batting averages for the current top 20 ranked batsmen in international Test cricket.

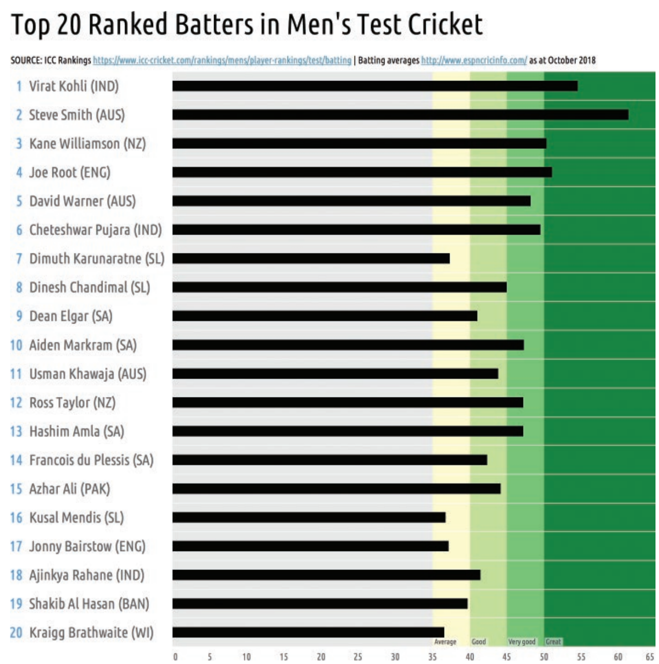


Figure 6.6 The Top 20 Ranked Batters in Men's Test Cricket (October 2018)

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. If you include axis-scale labels you should not need to label each bar value directly, as this will lead to label overload. Any colours used to indicate meaningful bandings or markers should be explained through the inclusion of a legend.

COMPOSITION: The bars should be proportionally sized according to the associated quantitative value – nothing more, nothing less – otherwise the perception of the bar sizes will be distorted. Most commonly, this means setting the quantitative value scale to an origin of zero. There is no significant difference in perception between vertically or horizontally arranged bullet charts; it will depend on which layout makes it easier to accommodate the range of values and to read the item labels associated with each bar. Aim to make the sorting of values in the chart as meaningful as possible.

VARIATIONS & ALTERNATIVES

Like the bar chart, bullet charts can also be used to show how values have changed over time. Further point markers (usually small circles or thin lines) can be included in the bullet chart to offer further useful comparisons and to optimise the interpretation.

WATERFALL CHART

ALSO KNOWN AS Cascade chart

C H R T S

COMPARISONS

REPRESENTATION DESCRIPTION

A waterfall chart provides details of how a total or net quantitative value has been formed through an ordered sequence of bars representing quantitative values for discrete categorical components. It is essentially a visual calculation showing different components of positive and negative values, represented by size and direction, to establish a running total. A common application of a waterfall chart would be to break down the calculation of profit as formed by different categories of income and expenditure.

EXAMPLE

Comparing the number of arriving and departing migrants from different regions that form the net level of migration in New Zealand.

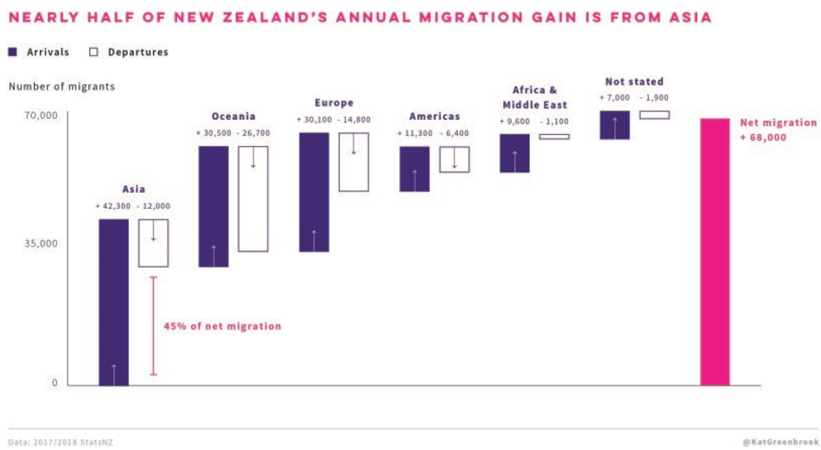


Figure 6.7 Nearly Half of New Zealand's Annual Migration Gain is From Asia, by Kat Greenbrook

PRESENTATION TIPS

ANNOTATION: Direct value labelling is usually applied to each step describing what each relates to as well as the quantitative amount. A dotted line is sometimes added to make more discernible what the running total is at each stage. Any colours used must be explained through the inclusion of a legend.

COLOUR: Attributes of colour are often established to classify visually each categorical stage or to distinguish further the positive and negative direction of quantitative values.

COMPOSITION: Most commonly a waterfall will be presented in landscape form with a left-to-right sequence arriving at a final total or net amount at the final right-side position.

VARIATIONS & ALTERNATIVES

One alternative would be to consider the stacked bar chart, as long as there were no negative quantitative values and all components are included that comprise a total, which represents a meaningful whole. The clustered bar chart may also be used to split the categorical parts of the final total in close proximity, but with all bars sized from a common baseline.

RADAR CHART

ALSO KNOWN AS Filled radar chart, star chart, spider diagram, parallel coordinates

C H R T S

COMPARISONS

REPRESENTATION DESCRIPTION

A radar chart plots values across multiple quantitative variables for one or several categorical items to enable general pattern forming. It uses a radial (circular) layout comprising several axes emerging from the centre-like spokes on a wheel, one for each variable. The quantitative values are then plotted along each scale using the attribute of position and then joined by connecting lines to form a unique geometric shape. Sometimes the lines or the shape fill is coloured for emphasis or for categorical differentiation when more than one item is plotted.

EXAMPLE

Comparing the global competitiveness scores across 12 'pillars' of performance for the UK versus Europe and North America.

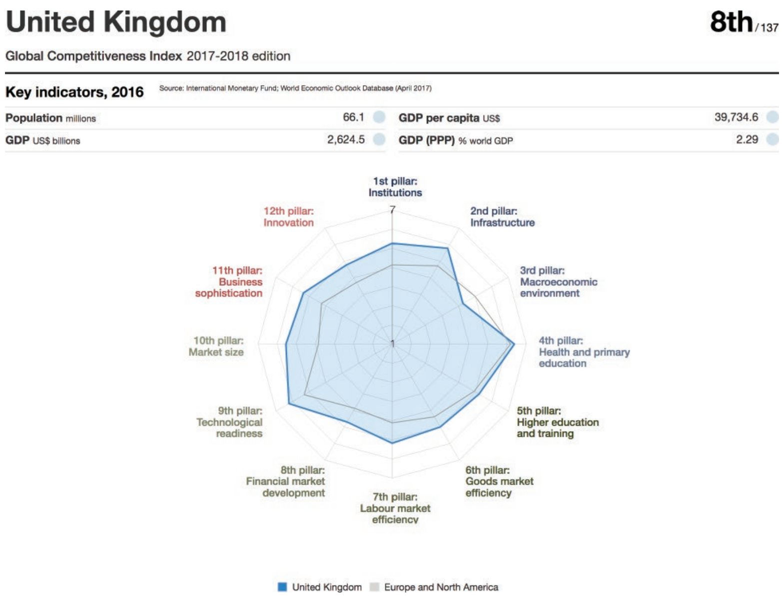


Figure 6.8 Global Competitiveness Report 2017–2018, by the World Economic Forum

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. Gridlines are only relevant if there are common scales across each quantitative variable. If so, the gridlines must be presented as straight lines, not concentric arcs, because the connecting lines joining up the values are themselves straight lines. If your quantitative values are on different scales, do not forget to display the values ranges on each. Any colours used must be explained through the inclusion of a legend.

COLOUR: When radar shapes are filled with a colour, sometimes a degree of transparency is applied to allow the chart apparatus to be still partially visible.

COMPOSITION: The cyclical ordering of the quantitative variables should be as meaningful as possible and consistent, as the shape formed will change for any ordering permutation. This will have a major impact on the readability and meaning of the resulting chart shape. A radar chart works best when the neighbouring pairings have some significant comparable value (such as values being plotted around the face of a clock or compass).

VARIATIONS & ALTERNATIVES

If you have common scales across the quantitative variables, a 'polar chart' is an alternative, should the radial layout be important to preserve. Otherwise, a 'bar chart' or 'dot plot' would be better options. While not strictly a variation, 'parallel coordinates' display a similar technique for plotting several independent quantitative measures in the same chart. The main difference is that parallel coordinates use a linear layout. If you have multiple category items, rather than plot them all on the same radar chart, consider using small multiples formed of distinct radars for each individual item instead.

POLAR CHART

ALSO KNOWN AS Coxcomb plot, polar area plot, circular barplot

C H R T S

COMPARISONS

REPRESENTATION DESCRIPTION

A radar chart plots values across multiple quantitative variables for one categorical item to enable general pattern forming. It uses a radial (circular) layout comprising several equal-angled circular sectors – like slices of a pizza, one for each variable. In contrast to the radar chart (which uses position along a scale), the polar chart uses variation in the size of the sector areas to represent values for each quantitative variable. It is, in essence, a radially arranged bar chart. Colour is an optional attribute, sometimes used to differentiate between different quantitative variables.

EXAMPLE

Comparing the capabilities of the four teams competing in Group F of the 2018 World Cup across seven distinct attributes.

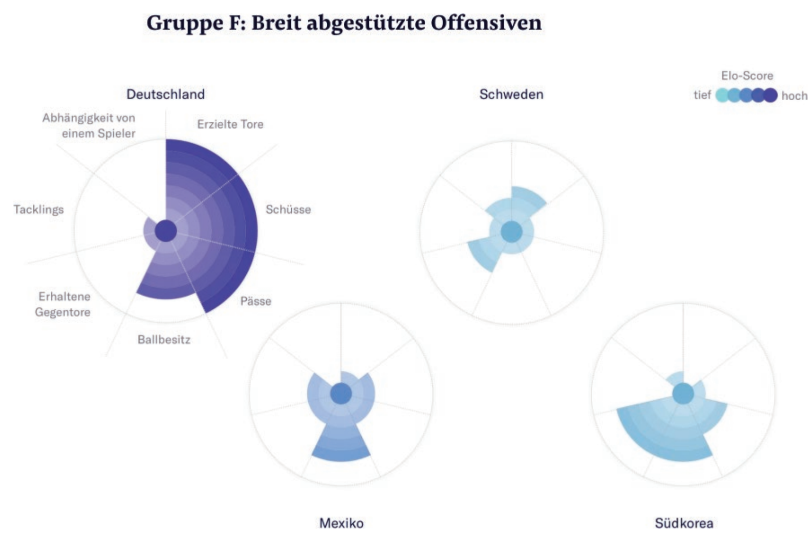


Figure 6.9 How Do National Teams Play? All 32 World Cup Participants in Direct Comparison [Translated], by NZZ Visuals

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. Gridlines are only relevant if there are common scales across each quantitative variable. If so, the gridlines must be presented as arcs reflecting the outer shape of each sector. Connecting lines joining up the values are themselves straight lines. Each sector typically uses the same quantitative scale for each quantitative measure but, on the occasions when this is not the case, do not forget to display the values ranges on each. Any colours used must be explained through the inclusion of a legend.

COMPOSITION: The cyclical ordering of the quantitative variables should be as meaningful as possible and consistent, as the shape formed will change for any ordering permutation. This will have a major impact on the readability and meaning of the resulting chart shape. A polar chart works best when the neighbouring pairings have some significant comparable value (such as values being plotted around the face of a clock or compass). The sizing of the sectors needs to be carefully calculated. Each sector should have a proportionally consistent angle of the whole and, to encode the quantitative values, the area of the sector, not the radius length, should be used.

VARIATIONS & ALTERNATIVES

If you have inconsistent scales across the quantitative variables, a ‘radar chart’ is an alternative should the radial layout be important to preserve. Otherwise, a ‘bar chart’ or ‘dot plot’ would be better options.

C H R T S

COMPARISONS

CONNECTED DOT PLOT

ALSO KNOWN AS Dot plot, dumbbell chart, range chart, dot chart, arrow chart

REPRESENTATION DESCRIPTION

A connected dot plot displays quantitative values for different primary category items with a secondary categorical breakdown enabling local comparisons. The plot is typically formed of two point marks plotting the quantitative value positions for each secondary categorical grouping. Joining the two points together is a connecting line which effectively represents the ‘delta’ (difference) between the two values through its size. Attributes of colour or variation in symbol are commonly used to distinguish the secondary categorical groupings.

EXAMPLE Comparing the typical salaries of women and men across a range of different job categories in the UK.

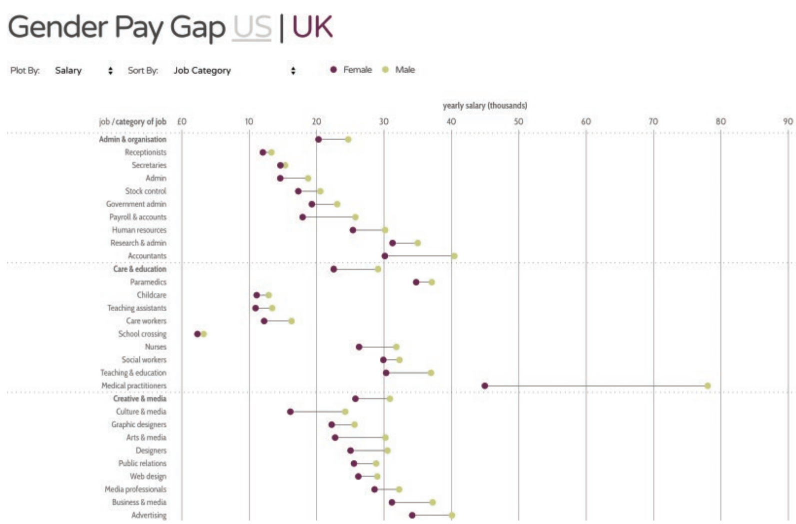


Figure 6.10 Gender Pay Gap UK, by David McCandless, Miriam Quick (Research) and Philippa Thomas (Design)

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. If you include axis-scale labels you should not need to label each value directly, as this will lead to label overload. Any colours used must be explained through the inclusion of a legend.

COLOUR: Colour may be used to help emphasise the directional basis of the connecting lines.

COMPOSITION: As the representation of the quantitative values is encoded through position along a scale and not size, the quantitative axis does not need to have a zero origin. However, a zero origin may be helpful to establish the scale of the differences depending on the subject matter being portrayed. If you do not commence from an origin of zero, this will need to be clearly annotated. Aim to make the sorting of values in the chart as meaningful as possible.

VARIATIONS & ALTERNATIVES

A variation in the application of the ‘connected dot plot’ would be to plot and compare values representative of two different points in time for the same measure. An alternative would be to use a variation of the ‘Gantt chart’, and rather than a single line starting from a minimum date and extending to a maximum date, you would just use this line to show the position and difference between quantitative values. An ‘arrow chart’ is an extension of this whereby the arrowhead is used to emphasise the directional basis of the line. Similarly, the ‘carrot chart’ uses line width tapering to indicate direction. If the number of secondary categories grows in number, the ‘dot plot’ would be useful to show the distribution of values rather than attempting to compare differences between just two values. A ‘clustered bar chart’ offers a further alternative for showing comparisons between secondary categorical dimensions.

PICTOGRAM

ALSO KNOWN AS Isotype chart, pictorial bar chart, symbol chart

REPRESENTATION DESCRIPTION

A pictogram displays quantitative values for different primary category items with the option for secondary categorical breakdown. The basis of the pictogram is the repetition in use of point marks, in the form of symbols or pictures, to represent an associated quantitative count. Each point mark may be representative of one or many quantitative units (e.g. a single symbol may represent 100 people). Secondary categorical dimensions can be incorporated through differentiation in the attribute of colour or symbol.

EXAMPLE
Comparing the grouped earning levels of the top 100 highest paid athletes during 2017.

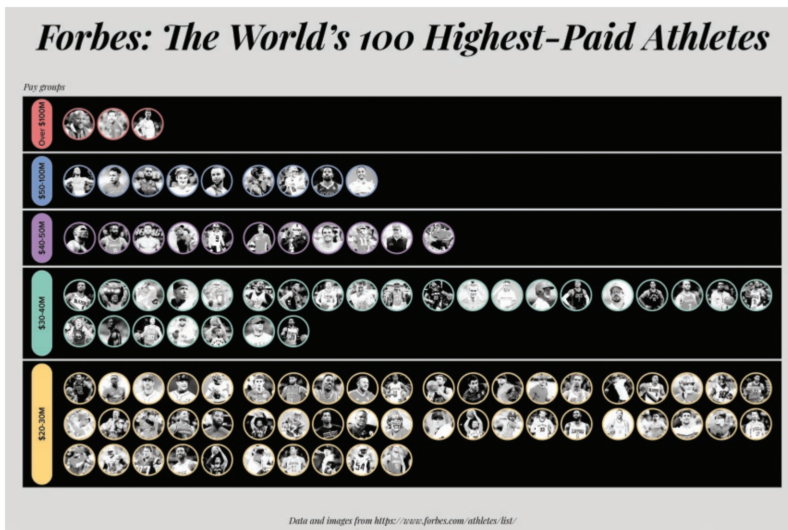


Figure 6.11 Forbes: The World's 100 Highest-paid Athletes, by Andy Kirk

PRESENTATION TIPS

ANNOTATION: The choice of symbols should be as recognisably intuitive as possible. If not, any legends should be presented close to the display to enable quick reference for determining the categorical and quantitative association of each symbol variation used.

COMPOSITION: If the quantities of markers exceed a single row, try to make the number of units per row logically 'countable', such as displaying in groups of 5, 10 or 100. To aid readability, make sure there is a sufficiently noticeable gap between clusters of grouped units. Aim to make the sorting of values in the chart as meaningful as possible.

VARIATIONS & ALTERNATIVES

When showing a part-to-whole relationship, the 'waffle chart' is similarly formed using point marks and symbol or colour attributes to differentiate the constituent parts of a whole.

PROPORTIONAL SYMBOL CHART

ALSO KNOWN AS Proportional shape chart, graduated symbol plot, bubble chart, circle packing diagram

C H R T S

COMPARISONS

REPRESENTATION DESCRIPTION

A proportional symbol chart displays quantitative values for different category items. The chart comprises shape marks with the size attribute (area) used to represent the quantitative value for each item. An attribute of colour may be used to accentuate the quantitative scale or organise marks by the distinct categories. Estimating and comparing the size of areas with accuracy is not as easy, so this chart type works best when you have a diverse range of quantitative value sizes.

EXAMPLE Comparing the market capitalisation ($) of companies involved in the legal sale of marijuana across different industry sectors.

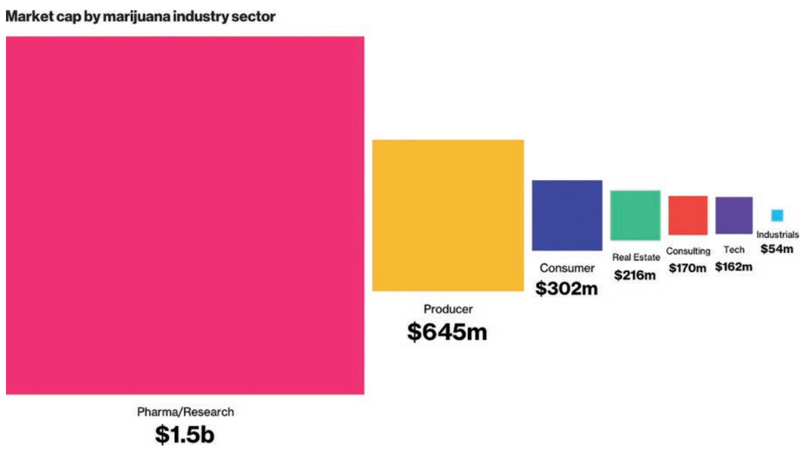


Figure 6.12 For These 55 Marijuana Companies, Every Day is 4/20, by Alex Tribou and Adam Pearce (Bloomberg Visual Data)

PRESENTATION TIPS

INTERACTIVITY: Proportional symbol charts may be accompanied by interactive features that let users select or mouseover individual shapes to reveal annotated values of the quantity and category.

ANNOTATION: If interactivity is not achievable, a quantitative size key should be included, or direct labelling incorporated. Though labelling can make a display cluttered (and be hard to fit when working with small-sized shapes) it will help overcome some of the limitations of judging area size. Any colours used must be explained through the inclusion of a legend.

COMPOSITION: The geometric accuracy of the shape mark size calculation is paramount: it is the area you are modifying, not the diameter/radius. Typically, the layout is quite free-form with no baseline or central gravity binding the display together. Otherwise, you might employ clustering or containers to help organise the categorical distinctions, though the colouring of each shape may already achieve this. Aim to make the sorting of the shapes in the chart as meaningful as possible.

VARIATIONS & ALTERNATIVES

Often, the data shown represents many parts of a whole. A 'circle packing diagram' uses circular shapes and packs the contents into a neat circular layout representing a whole. The 'bubble plot' also uses differently sized shapes (usually circles) but the position is meaningful across two quantitative variable dimensions. By removing the size attribute (and effectively replacing the shape mark with a point mark) you could use the quantity of points clustered together for different categorical totals to create a variation of the 'pictogram'.



WORD CLOUD

ALSO KNOWN AS Tag cloud, proportional symbol chart

REPRESENTATION DESCRIPTION

A word cloud shows the frequency of individual word items within a passage of textual data. Each item is represented by words and then the font size of each is scaled according to the frequency of its usage. Words already have varied lengths, so it is important to remember that it is effectively the area of the word, not its length, that encodes its quantitative measure.

EXAMPLE
Comparing the frequency of words used in Chapter 1 in the first edition of this book.

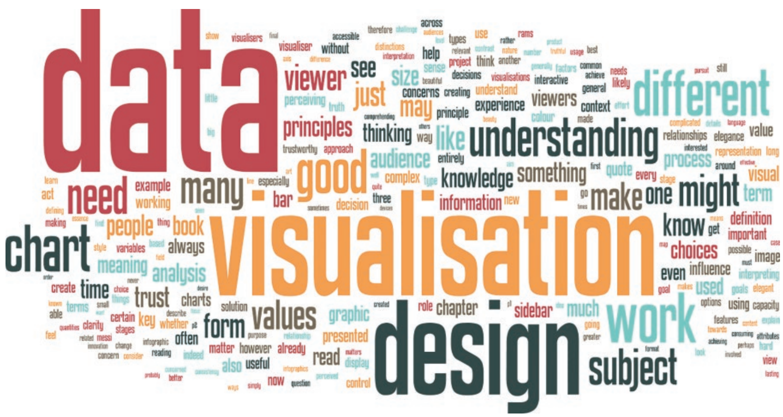


Figure 6.13 Comparing the Frequency of Words Used in Chapter 1 in the First Edition of this Book

PRESENTATION TIPS

INTERACTIVITY: Interactivity that lets users interrogate, filter and scrutinise the words in more depth, perhaps presenting examples of their usage in a passage, can be quite useful features to enhance the value of a word cloud.

ANNOTATION: Word clouds are most useful when you are trying to form a quick sense of some of the dominant keywords used in the text. Relative comparisons can be aided by including a key to explain how the font size scales equate to word frequency. Any colours used must be explained through the inclusion of a legend.

COMPOSITION: The arrangement of the words within a word cloud is typically based on a layout process that calculates the best placement of each word to occupy the optimum space.

VARIATIONS & ALTERNATIVES

Variations may include colours being used as a second form of quantitative encoding to accentuate the larger frequencies further or to organise useful groupings categorically. You might also consider using containers to separate out different clusters. Any alternative method from this categorical family of charts would more usefully display the counts of text, such as a bar chart or a proportional shape chart where the word label sits inside a sized shape mark.

HEAT MAP

ALSO KNOWN AS Matrix chart, mosaic plot, table chart, XY heatmap, 2D density plot

REPRESENTATION DESCRIPTION

A heat map displays quantitative values across the intersections of two categorical and/or discrete quantitative dimensions. The chart comprises two categorical axes with each distinct value presented across the row and column headers of a tabular layout. The corresponding cells effectively house a point mark with the attribute of colour (usually, colour lightness) used to represent the associated quantitative value.

EXAMPLE Comparing the average number of daily births across England and Wales between 1995 and 2014.

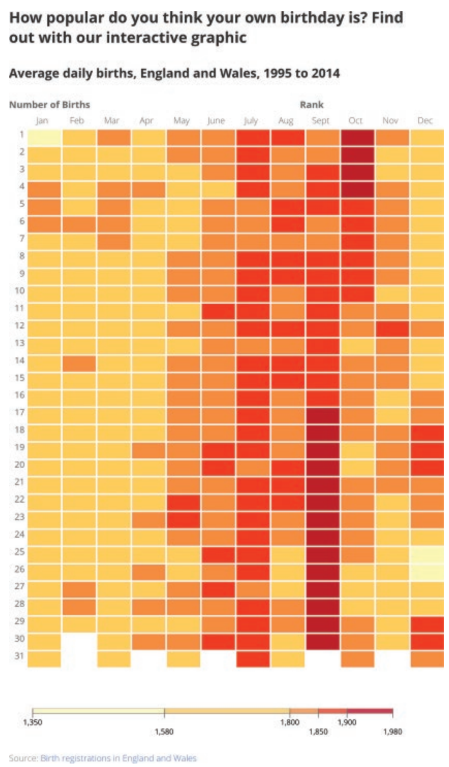


Figure 6.14 How Popular is Your Birthday?, by ONS Digital Content team

VARIATIONS & ALTERNATIVES

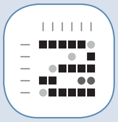
A 'radial heat map' offers a structure variation whereby the table may be portrayed using a circular layout. As with any radial display, this is really of value only if the cyclical ordering means something for the subject matter. A variation would see the colour lightness replaced by a categorical colouring approach if the values plotted were not quantitative in nature. An alternative chart approach would be the 'matrix chart' using the size of a shape or the frequency of clustered point marks to indicate a quantitative value.

PRESENTATION TIPS

ANNOTATION: Direct value labelling is possible but normally a clear legend to indicate colour associations will suffice. It is not easy for the eye to determine the exact quantitative values represented by the colours, even if there is a colour scale provided; heat maps mainly facilitate more a gist of the order of magnitude.

COLOUR: Decisions need to be made about whether to use a smooth colour gradient or employ discrete classifications for different value intervals. Different approaches will affect the patterns that emerge. There is no single right answer – you will arrive at it largely through trial and error/experimentation – but it is important to consider, especially when you have a diverse distribution of values.

COMPOSITION: Logical sorting and/or even grouping of the categorical values along each axis will aid readability and may help to surface key relationships.



MATRIX CHART

ALSO KNOWN AS Table chart, correlogram

C H R T S

COMPARISONS

REPRESENTATION DESCRIPTION

A matrix chart displays quantitative values across the intersections of two categorical and/or discrete quantitative dimensions. The chart comprises two categorical axes with each distinct value presented across the row and column headers of a tabular layout. The corresponding cells effectively house a geometric shape with scaled area size or clusters of point marks repeated in quantity to represent the associated quantitative value. Attributes of colour are often used visually to distinguish further categorical detail.

EXAMPLE
Comparing the number of Nobel Laureates by award category and country of birth.

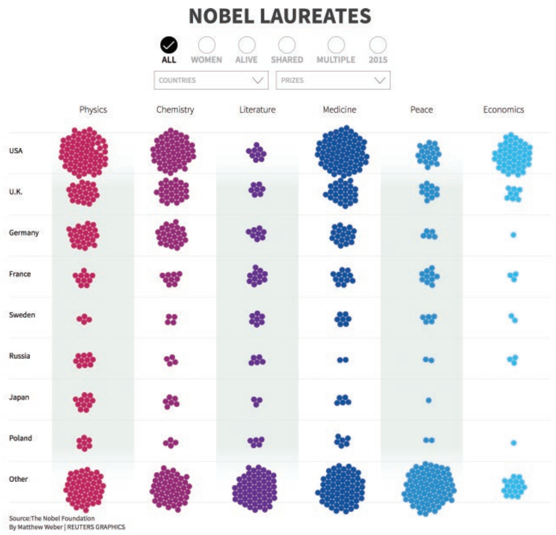


Figure 6.15 Nobel Laureates, by Matthew Weber (Reuters Graphics)

PRESENTATION TIPS

INTERACTIVITY: When using point mark clusters, interactive features can be useful to enable users to discover the labels of each item through tooltips.

ANNOTATION: When shape marks are used, direct value labelling is possible but normally a clear key to indicate the size associations will suffice. Any colours used must be explained through the inclusion of a legend.

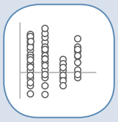
COLOUR: Colours may not be necessary because the tabular layout already establishes separation across the two categorical dimensions. However, employing an additional attribute of colour can help to distinguish further the horizontal or vertical categorical values.

COMPOSITION: If there are diverse value sizes with some especially large outliers, it may be necessary for the size of the shape marks or the quantity of point clusters to outgrow the space of the relevant cell. This might help to emphasise editorially the outlier status. Controlling this may not be possible, in which case the largest quantitative value will usually fill no more than the maximum space available. Logical sorting (and maybe even sub-grouping) of the categorical values along each axis will aid readability and may help surface key relationships. The geometric accuracy of the shape mark size calculation is paramount: it is the area you are modifying, not the diameter/radius.

VARIATIONS & ALTERNATIVES

A variation may involve the intersecting cells being representative of categorical values (nominal or ordinal), and therefore you might substitute quantitative attributes of size or quantity with variation in symbols and/or colour attributes.

An alternative chart type might be the 'heat map', which similarly indicates quantitative values at the intersections of two categorical and/or discrete quantitative dimensions.



DOT PLOT

ALSO KNOWN AS Univariate scatter plot, 1D scatter plot, instance chart, strip plot, barcode chart

REPRESENTATION DESCRIPTION

A dot plot displays the distribution of quantitative values for data items, sometimes broken down by a categorical dimension, to show the range and shape of quantities. The plot is typically formed of point marks positioned along a quantitative scale. The point marks may be small circles or thin lines ('strips'). If categorical differentiation is necessary, attributes of colour or variation in symbol may be employed within a single plot, otherwise several separate plot views will be created for each discrete category grouping.

EXAMPLE Comparing the ranking distribution of the top 200 billionaires by industry.

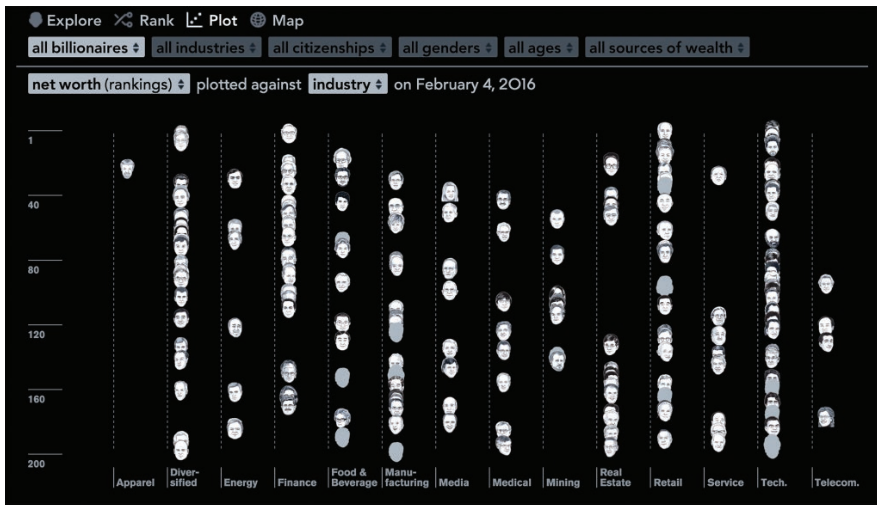


Figure 6.16 Bloomberg Billionaires, by Bloomberg Visual Data (Design and Development), Lina Chen and Anita Rundles (Illustration)

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. If you include axis-scale labels you should not need to label each value directly, as this will lead to label overload. Direct labelling will normally be restricted to noteworthy points only. Any colours used must be explained through the inclusion of a legend.

COLOUR: Colour may be used to establish the focus of certain points and/or distinction between different sub-category groups to assist with interpretation. To overcome occlusion caused by plotting several marks at the same value position, you might use unfilled or semi-transparent filled circles to convey value frequency.

COMPOSITION: As the representation of the quantitative values is encoded through position along a scale, the quantitative axis does not need to have a zero origin, unless this is meaningful to the subject. If you do not commence from an origin of zero, this will need to be clearly annotated.

VARIATIONS & ALTERNATIVES

A variation in the encoding of the dot plot may see the point marks replaced by shape marks (usually circles) in order to represent a second quantitative measure through size variation. This might be a useful method to represent the frequency of observations when several items share a similar value. The variation in the role of the dot plot would be through the 'instance chart', which plots events over a temporal axis rather than a quantitative scale. An alternative chart type would be the 'beeswarm plot', especially when you have a non-uniform distribution of values that cluster around similar quantities. You could also use a 'scatter plot' with its second axis offering the scope to plot two data quantitative variables with the items spread across the associated coordinate positions.

BEESWARM PLOT

ALSO KNOWN AS Jitter plot

C H R T S

DISTRIBUTIONS

REPRESENTATION DESCRIPTION

A beeswarm plot displays the distribution of quantitative values for data items to show the range and shape of quantities. The plot is typically formed of point marks, usually small circles, positioned along a quantitative scale. The points are then evenly distributed using a second dimension of space above and below the quantitative axis baseline, not to represent any quantitative measure, but to accommodate closely packed points that have similar value positions. If categorical differentiation is necessary, attributes of colour or variation in symbol may be employed within a single plot, otherwise several separate plot views will be created for each discrete category grouping.

EXAMPLE

Comparing the distribution of household incomes for a simulated population of Chicago residents broken down by ethnic group.

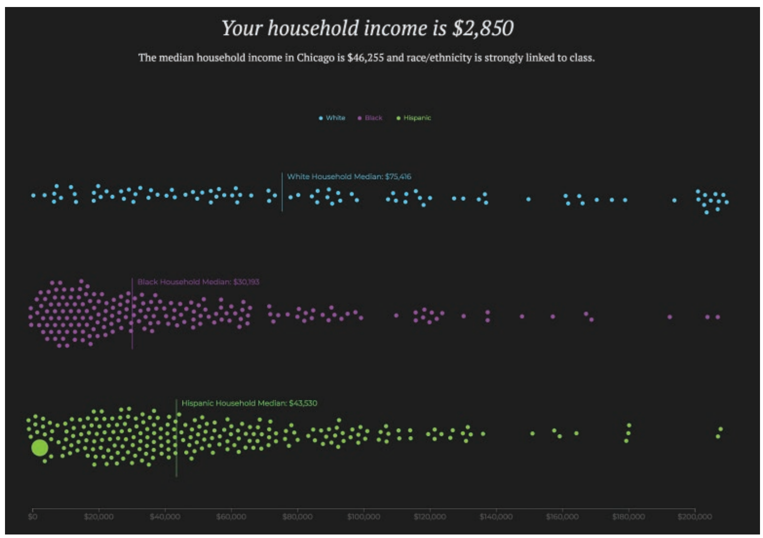


Figure 6.17 Is Your Child Ready for School?, by Gabrielle LaMarr LeMee

PRESENTATION TIPS

INTERACTIVITY: Interactive features can be useful to enable users to discover the value labels of each item through tooltips.

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. If you include axis-scale labels you should not need to label each value directly, as this will lead to label overload. Direct labelling will normally be restricted to noteworthy points only. Any colours used must be explained through the inclusion of a legend.

COLOUR: Colour may be used to establish the focus of certain points and/or distinction between different sub-category groups to assist with interpretation.

COMPOSITION: As the representation of the quantitative values is encoded through position along a scale, the quantitative axis does not need to have a zero origin, unless this is meaningful to the subject. If you do not commence from an origin of zero, this will need to be clearly annotated.

VARIATIONS & ALTERNATIVES

A variation in the encoding of the beeswarm plot may see the point marks replaced by shape marks (usually circles) in order to represent a second quantitative measure through size variation. An alternative chart type would be the 'dot plot', which removes the second dimension spread of values and overlays similar values. You could also use a 'histogram' to show the frequency and distribution of values in discrete quantitative groupings.

HISTOGRAM

ALSO KNOWN AS Bar chart (wrongly), population pyramid

C H R T S

DISTRIBUTIONS

REPRESENTATION DESCRIPTION

A histogram displays the frequency and distribution of quantitative measurements across grouped values for data items. Whereas bar charts compare quantities for discrete nominal categories, a histogram uses discrete quantitative 'bins' to form ordinal value groupings. The representation is formed using variation of line size (if the value groupings have equal intervals) or of shape area (if the value groupings have unequal value intervals) to represent the frequency of measurements.

EXAMPLE Comparing the distribution of lightness range among common foundation products sold in four regions.

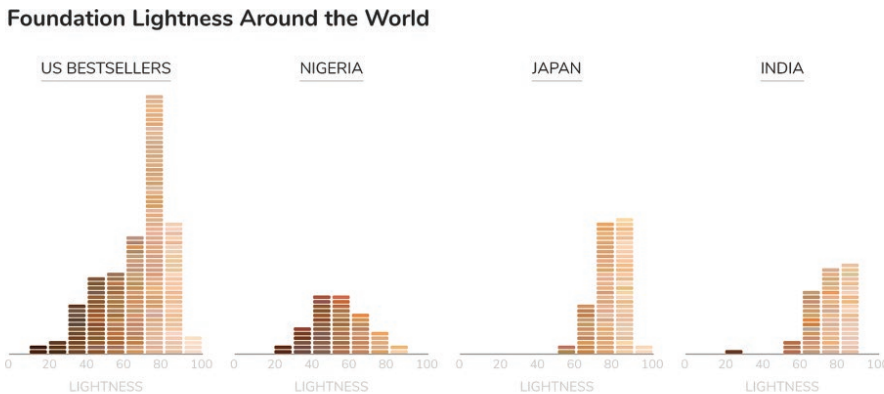


Figure 6.18 Beauty Brawl: How Inclusive are Beauty Brands Around the World?, by Amber Thomas, Jason Li and Divya Manian for 'The Pudding'

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. If you include axis-scale labels you should not need to label each value directly, as this will lead to label overload.

COMPOSITION: Unlike the bar chart there should be no (or, at most, a very thin) gap between bars to help the collective shape of the frequencies emerge. The sorting of the quantitative value bins must be presented in ascending order so that the reading of the overall shape preserves its meaning. The number of value bins and the range of values covered by each have a prominent influence over the appearance of the histogram and the usefulness of what it might reveal: too few bins may disguise interesting nuances, patterns and outliers; too many bins and the most interesting shapes may be abstracted by noise above signal. There is no singular best approach, the right choice simply arrives through experimentation and iteration.

VARIATIONS & ALTERNATIVES

For an analysis that looks at the distribution of values across two dimensions, such as populations by age group across binary gender categories, you might consider a 'back-to-back histogram' also commonly known as a 'population pyramid'. A 'box-and-whisker plot' is an alternative approach that reduces the display of the distribution of values to just five key statistical measures. To reveal more granular detail, the 'dot plot' and 'beeswarm plot' display all items individually across a quantitative scale.

C H R T S

DISTRIBUTIONS

DENSITY PLOT

ALSO KNOWN AS Ridgeline plot

REPRESENTATION DESCRIPTION

Density plots display the frequency and distribution of quantitative values for data items. Whereas histograms compare quantities using discrete quantitative 'bins' to form ordinal value groupings, a density plot can be considered a smoothed histogram. The plot is typically formed of a quantitative scale along which a line connects measurements of the frequency of each quantitative value. The line gets higher as the frequency gets higher. The connected line is then smoothed using various statistical techniques (that will depend on the subject context) and the area below is filled with colour to help visibility of the resulting shape. This creates the appearance of an 'area chart'. Often the density plot comprises multiple rows to separate observations across discrete category groupings.

EXAMPLE

Comparing the distribution of scores allocated to a selection of words or phrases indicating the perceived level of positivity or negativity.

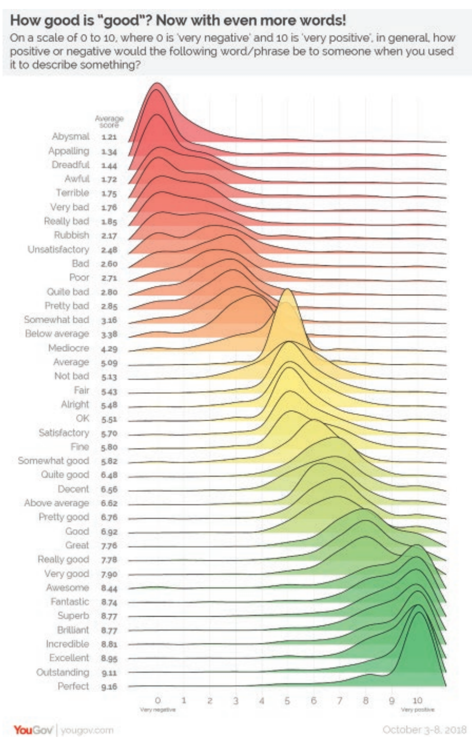


Figure 6.19 How Good is 'Good'?, by Matthew Smith

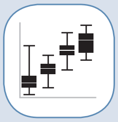
VARIATIONS & ALTERNATIVES

A variation in the design of a density plot is the 'violin plot' whereby the shape of distribution is plotted symmetrically creating a two-sided violin-like, rather than the one-sided shape of the density plot. An alternative role for the density plot would be in the form of an 'area chart', which plots quantitative trends over a temporal axis rather than a quantitative scale. An alternative chart type would be the 'beeswarm plot' to show the quantitative values of individual data items or a 'histogram' to show the frequency and distribution of values in discrete quantitative groupings.

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines is not usually necessary with density plots as they are more about getting a sense of the shape and patterns.

COMPOSITION: Depending on the nature of the quantitative measurements, and in particular the presence of outlier shapes in the distribution of values, the density plot is often presented in a way whereby high-value area 'spikes' intrude into and over the row space occupied by categories above. The arrangement of discrete categories is important to avoid too much occlusion and/or wasted empty space.



BOX-AND-WHISKER PLOT

ALSO KNOWN AS Box plot, candlestick chart, OHLC chart

C H R T S

DISTRIBUTIONS

REPRESENTATION DESCRIPTION

A box-and-whisker plot displays the distribution and shape of a series of quantitative values for different categories. The display is formed by a combination of lines and point markers to indicate (through position and length), typically, five different statistical measures. Three of the statistical values are common to all plots: the first quartile (25th percentile), the second quartile (or median) and the third quartile (75th percentile) values. These are displayed with a box (effectively a wide bar) positioned and sized according to the first and third quartile values with a marker indicating the median. The remaining two statistical values vary in definition: usually the minimum and maximum values or the 10th and 90th percentiles. These statistical values are represented by extending a line beyond the bottom and top of the main box to join with a point marker indicating the appropriate position. These are the whiskers. A single plot will be produced for each relevant, discrete category grouping.

EXAMPLE Comparing the distribution of annual earnings 10 years after starting school for graduates across the eight Ivy League colleges.

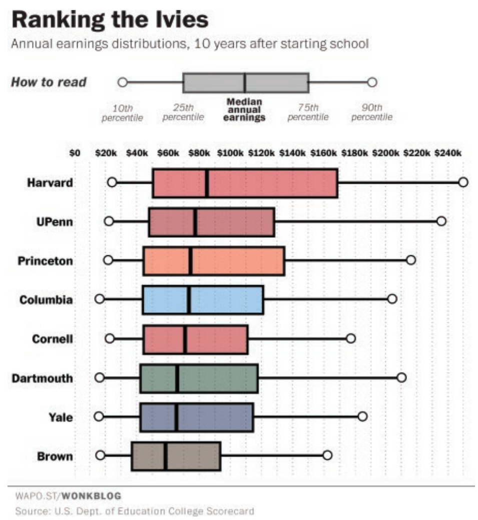


Figure 6.20 This Chart Shows How Much More Ivy League Grads Make Than You, by Christopher Ingraham (*Washington Post*)

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. If you include axis-scale labels you should not need to label each value directly, as this will lead to label overload. Direct labelling will normally be restricted to noteworthy points only.

COMPOSITION: The quantitative value axis does not need to commence from zero, unless it means something significant to the interpretation, as the ranges of values themselves do not necessarily start from zero and the focus is more on the statistical properties between the outer values. There is no significant difference in perception between vertically or horizontally arranged box-and-whisker plots; it will depend on which layout makes it easier to accommodate the range of values and to read the item labels associated with each bar. When you have several plots in the same chart, where possible try to make the categorical sorting meaningful, perhaps by organising values in ascending or descending order based on the median value.

VARIATIONS & ALTERNATIVES

Variations mainly concern changing the number of statistical measures included in the display. Sometimes you might remove the ‘whiskers’ to show just the 25th and 75th percentiles through the lower and upper parts of the ‘box’. The ‘candlestick chart’ (or ‘OHLC chart’ used in stock market analysis to track the opening, highest, lowest and closing prices of stocks) uses a similar method and is often used to show the distribution and milestone quantitative values for events that encounter constant change, such as stock market analysis over a given time frame based on showing the opening, highest, lowest and closing prices.

PIE CHART

ALSO KNOWN AS Pizza chart, donut chart (wrongly)

C H R T S

PART-TO-WHOLE

REPRESENTATION DESCRIPTION

A pie chart shows how proportions of quantities for different constituent categories make up a whole. It uses a circular display divided into sectors for each category, with the angle representing the percentage proportions and attributes of colour to separate the discrete categories. The resulting size of the sector (in area terms) is a spatial by-product of the angle and so offers an additional means for judging values. The total of all sector values must be 100% and the constituent parts must be exclusive and representative of a meaningful 'whole', otherwise the chart will be corrupted.

EXAMPLE Comparing the proportion of Michael Schumacher's F1 races by result group.

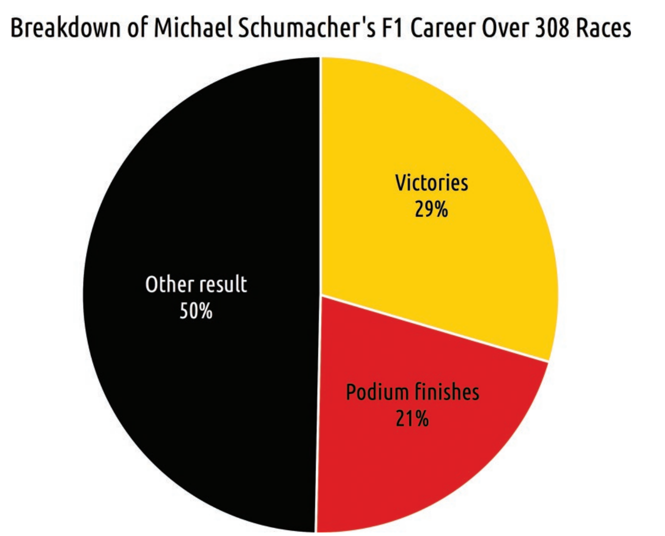


Figure 6.21 Breakdown of Michael Schumacher's F1 Career Over 308 Races

PRESENTATION TIPS

ANNOTATION: Directly labelling each category and associated value can enhance readability but may create inelegant clutter depending on the shape of the data and the size of the label values. Any colours used must be explained through the inclusion of a legend.

COLOUR: Colour is used to classify the categorical associations of each sector, so aim to vary the hue property of each colour to maximise the visible difference. When you have multiple sectors, you might choose to emphasise only two or three parts through editorial selection.

COMPOSITION: Positioning the first slice at the vertical 12 o'clock position gives a useful baseline to help judge the first sector angle value. The ordering of sectors using descending values or ordinal characteristics helps with the overall readability and allocation of effort.

VARIATIONS & ALTERNATIVES

The principal variation of the pie chart would be the 'donut chart'. Its function is exactly the same, but the donut has the centre removed, often to accommodate a labelling property. This removes the possibility of judging the sector angles at the circle origin, so the reading is formed by the arc lengths. The role of a pie chart is primarily about being able to compare a 'part to a whole' than being able to compare one part to another part. If you want to display and compare multiple parts, the 'bar chart' will offer a better option. For showing many parts, especially if they are organised into hierarchical groupings, the 'treemap' is a good option. Depending on the allocated space, a 'stacked bar chart' may provide an alternative layout to the pie chart, especially if your categorical values have an ordinal relationship. A 'nested shape chart', typically based on square or circle marks, enables comparisons across a series of one-part-to-whole relationships showing absolute values (through size) and proportions (through relative size).

WAFFLE CHART

ALSO KNOWN AS Square pie, unit chart, grid plot

REPRESENTATION DESCRIPTION

A waffle chart shows how proportions of quantities for different constituent categories make up a whole. It uses a square display divided typically into 100 points arranged in a grid layout. Each constituent proportion is displayed through colour coding the relevant number of points. The role of the waffle chart is to simplify the counting of proportions in contrast to the angle judgements of the pie chart, though the display is limited to only showing integer values. The total of all sector values must be 100% and the constituent parts must be exclusive and representative of a meaningful 'whole', otherwise the chart will be corrupted.

EXAMPLE Comparing the percentage of hours of sun by month during 2017 in Nuorgam, Finland.

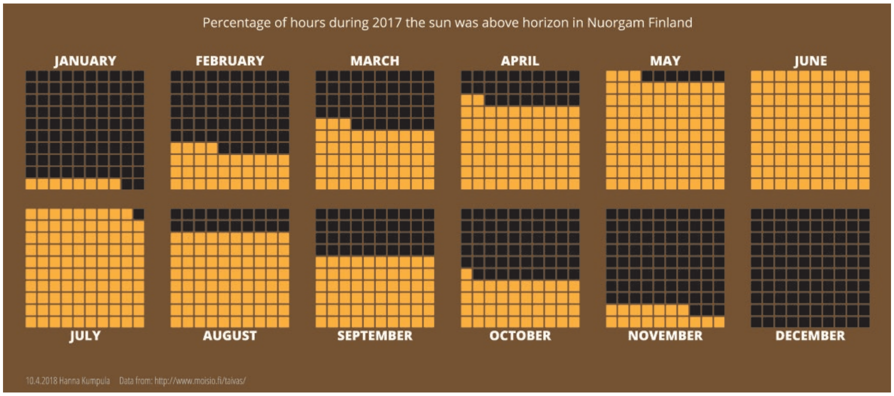


Figure 6.22 Percentage of Hours During 2017 the Sun was Above the Horizon in Nuorgam, Finland, by Hanna Kumpula (@kumpulahanna)

PRESENTATION TIPS

ANNOTATION: Chart apparatus is rarely applied to a waffle chart, though direct labelling may be included, perhaps using a nearby caption to indicate a category and quantitative label. Any colours used must be explained through the inclusion of a legend.

COLOUR: Adding outlines to each point mark (grid cell or circle) can be useful to help discern individual units.

COMPOSITION: A waffle chart is quicker to read when clusters of units, such as groups of five or ten, can be easily recognised. You may therefore seek to arrange the cells in groups to facilitate this. When you have several parts in the same waffle chart, where possible try to make the categorical sorting meaningful.

VARIATIONS & ALTERNATIVES

Rather than using colour, sometimes variations in symbols will be used to classify different categories or groupings. For example, you might see figures or gender icons used to show the makeup of a given sample population. A variation in the role of a waffle chart is to show quantitative counts rather than proportions of a whole, and this approach somewhat overlaps with applications of the 'pictogram'. A 'nested shape chart' using sized rectangular shapes may provide an alternative way of showing a part-to-whole relationship while also occupying a squarified layout.

STACKED BAR CHART

ALSO KNOWN AS Stacked chart, packed bars

C H R T S

PART-TO-WHOLE

REPRESENTATION DESCRIPTION

A stacked bar chart shows how quantitative values for different constituent categories make up a whole across major category items. The proportion of each constituent categorical 'part' is represented by separate bars that are sized according to their quantitative proportion and then stacked to create the whole. Sometimes the whole is standardised to represent 100%, otherwise it will be representative of an absolute total. Colour attributes are used to classify the discrete categorical parts. Stacked bar charts often work best when the categories are ordinal in nature, and it is the overall pattern of spread across the whole that is important. If the parts are representative of nominal categories, judging and comparing the size of individual stacked parts become quite hard without a common baseline, so you might seek to reduce the number of discrete values.

EXAMPLE Comparing the degree of trust held by Australians for different institutions.

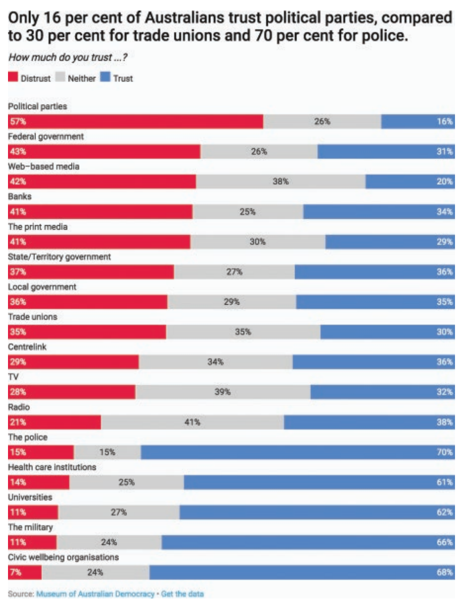


Figure 6.23 In a Nation of Cynics, We're Flocking to the Fringe, by ABC

VARIATIONS & ALTERNATIVES

The main alternative would be to create multiples of bar charts each showing the quantitative values for just a single constituent part for each major category item. The 'waterfall chart' splits out the individual constituent parts to create a step-by-step breakdown of a single stacked whole. Like their unstacked siblings, stacked bar charts can also be used to show how value proportions have changed over time.

PRESENTATION TIPS

ANNOTATION: Direct value labelling can become very cluttered when there are many parts, so you may choose to focus only on labelling noteworthy values. Axis scales using logical intervals will be helpful, as will the inclusion of gridlines, especially highlighting key features such as the 50% position when your data is displaying a 100% stacked total. Any colours used must be explained through the inclusion of a legend.

COLOUR: If you are representing categorical ordinal data, colour can be astutely deployed to give a sense of the general balance of values within the whole, but this will only work if their sorting arrangement within the stack is logically applied. For categorical nominal data, ensure the stacked parts have sufficiently different colour hues so their distinct bar lengths can be easily detected.

COMPOSITION: The bars should be proportionally sized according to the associated quantitative value – nothing more, nothing less – otherwise the perception of the bar sizes will be distorted. Most commonly, this means setting the quantitative value scale to an origin of zero. There is no significant difference in perception between vertically or horizontally arranged stacked bar charts; it will depend on which layout makes it easier to accommodate the range of values and to read the item labels associated with each cluster. Including a noticeable gap between each stack of bars will help to preserve a clear distinction between the primary category items. Aim to make the sorting of values in the chart as meaningful as possible.

DIVERGING BAR CHART

ALSO KNOWN AS Back-to-back bar chart, paired bar chart, spine chart

C **H** R T S

PART-TO-WHOLE

REPRESENTATION DESCRIPTION

A diverging bar chart shows how quantitative values for different constituent categories make up a whole across major category items. The proportion of each constituent categorical 'part' is represented by separate bars that are sized according to their quantitative proportion and then stacked to create the whole. Sometimes the whole is standardised to represent 100%, otherwise it will be representative of an absolute total. In contrast to the stacked bar chart, the diverging bar chart arranges constituent categorical parts either side of a common baseline depending on the discrete nominal or ordinal relationships that benefit from such separation. Colour attributes are commonly used to classify the discrete categorical parts.

EXAMPLE Comparing the responses to a survey question asking for opinions about the legality of abortion across different demographic categories.

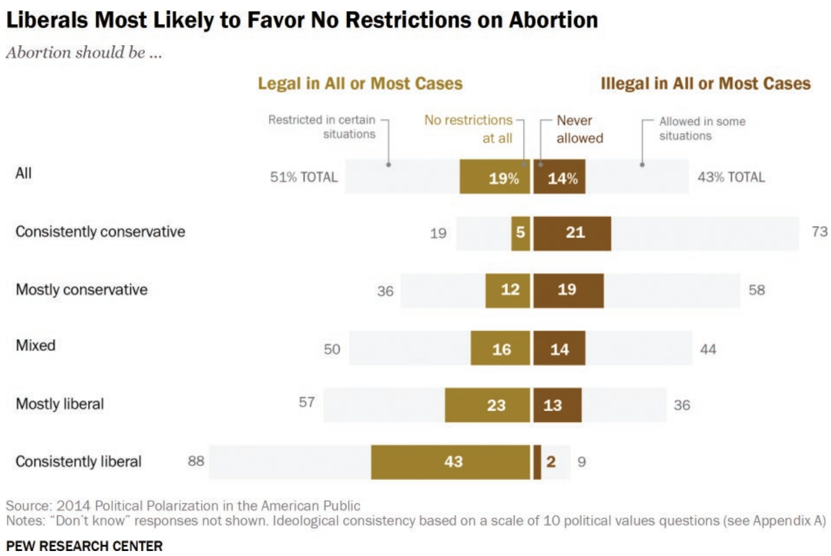


Figure 6.24 Political Polarization in the American Public, Pew Research Center, Washington, DC (February, 2015) (http://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/)

PRESENTATION TIPS

ANNOTATION: Direct value labelling can become cluttered when there are many constituent parts, so you may choose to focus only on labelling noteworthy values. Axis scales using logical intervals will be helpful, as will the inclusion of gridlines. Any colours used must be explained through the inclusion of a legend.

COLOUR: If you are representing categorical ordinal data, colour can be astutely deployed to give a sense of the general balance of values within the whole, but this will only work if their sorting arrangement within the stack is logically applied. For categorical nominal data, ensure the stacked parts have sufficiently different colour hues so their distinct bar lengths can be easily detected.

COMPOSITION: There is no significant difference in perception between vertically or horizontally arranged diverging bar charts; it will depend on which layout makes it easier to accommodate the range of values and to read the item labels associated with each cluster. Including a noticeable gap between each stack of bars will help to preserve a clear distinction between the primary category item. Aim to make the sorting of values in the chart as meaningful as possible.

VARIATIONS & ALTERNATIVES

A variation would be a 'diverging histogram' whereby the major categories have ordinal qualities, like increasing age groups, and the resulting shape of the chart has meaning about the distribution of values. If you want to facilitate direct comparison, a 'clustered bar chart' showing adjacent bars may offer a better alternative solution.

MARIMEKKO CHART

ALSO KNOWN AS Mekko chart, mosaic plot, proportional stacked bar

C H R T S

PART-TO-WHOLE

REPRESENTATION DESCRIPTION

A Marimekko chart is effectively a two-dimensional stacked bar chart with variation in size for both height and width to display parts of a whole simultaneously across two dimensions. It is often used to contextualise percentage part-to-whole comparisons of major categories with a second dimension of absolute numbers that make up a total. Attributes of colour are commonly used to provide categorical classifications.

EXAMPLE
Comparing the proportion and number of competitors by gender across all Summer Olympic Games.

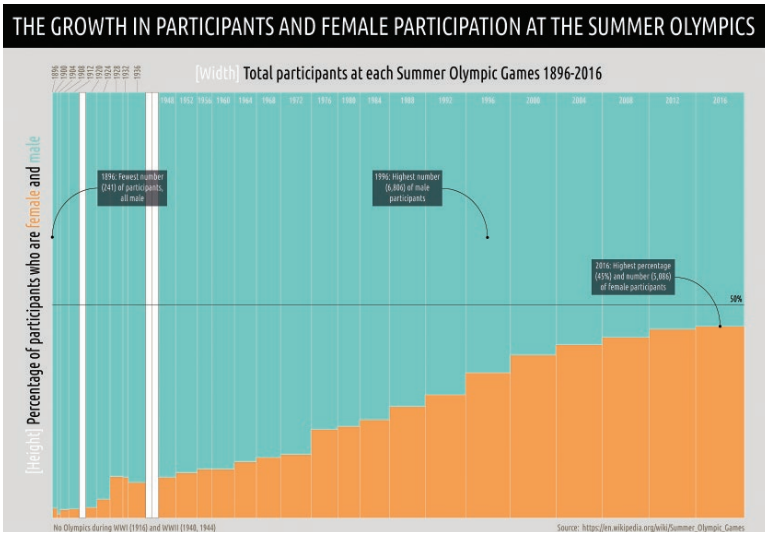


Figure 6.25 The Growth in Participants and Female Participation at the Summer Olympics

PRESENTATION TIPS

ANNOTATION: With two quantitative axes and two dimensions of categorical division, labelling Marimekko charts can become quite cluttered. At the very least, the two axes should be clearly titled, and some size scales provided, either through axis interval labelling or direct labelling of noteworthy items. Any colours used must be explained through the inclusion of a legend.

COLOUR: It will usually be possible to distinguish classifications visually across only one of the categorical dimensions.

VARIATIONS & ALTERNATIVES

An alternative to the Marimekko chart would be the treemap which shows part-to-whole relationships when there are many category parts and there is some hierarchical organisation of those categories.

TREEMAP

ALSO KNOWN AS Heat map (wrongly)

C H R T S

HIERARCHIES

REPRESENTATION DESCRIPTION

A treemap is an enclosure diagram providing a hierarchical display that shows how quantitative values for different constituent categorical parts make up a whole. It uses a contained rectangular layout, often termed **squarified**, representing the 100% total. This is divided into proportionally sized shape marks (rectangular tiles) for the quantitative values associated with each categorical part. The organisation of each shape is based on a tiling algorithm to optimise the overall space usage and to cluster related categories into larger rectangle-grouped containers. Attributes of colour are often used to represent further quantitative measures or categorical associations. Treemaps are most commonly used, and of most value, when there are many parts to the whole. The constituent parts must be exclusive and the total representative of a meaningful 'whole', otherwise the chart will be corrupted.

EXAMPLE Comparing the relative value and daily changes of market capital for stocks across the S&P 500 index grouped by sectors and industries.



Figure 6.26 Finviz: Standard and Poor's 500 Index Stocks (www.finviz.com)

PRESENTATION TIPS

INTERACTIVITY: Interactive features that enable selection events to trigger annotated tooltips can be useful, providing direct value labels and details. There may also be scope for modifications to temporal dimensions, changing the sizes and colouring accordingly, or zooming techniques to get a closer view of small constituent parts. **ANNOTATION:** Group or container labels can be hard to allocate space to efficiently, so borders are usually applied to indicate the relevant enclosure areas. Effective direct value labelling becomes difficult as the rectangles get smaller, so only the most prominent values may be annotated, especially if interactivity is not available. Any colours used must be explained through the inclusion of a legend. **COMPOSITION:** As the tiling algorithm used by any given tool to create a treemap will be focused on optimising the dimensions and arrangement of the rectangular shapes, treemaps may not always be able to facilitate meaningful sorting of high to low values within each enclosure. However, you will generally find larger areas appear in the top left and work outwards towards the smaller constituent parts.

VARIATIONS & ALTERNATIVES

A variation of the treemap sees the overall rectangular layout replaced by a circular one and the tiles represented by polygonal shapes. These are known as 'Voronoi treemaps' as the tiling algorithm is informed by a Voronoi tessellation. The 'circle packing diagram', a variation of the 'bubble chart', similarly shows many parts to a whole but uses non-tessellating circular shapes. The 'Marimekko chart' is similar in appearance to a treemap but, in contrast to the treemap's hierarchical display, presents a breakdown of quantitative percentages and/or absolute values across two categorical dimensions.

SUNBURST CHART

ALSO KNOWN AS Icicle chart, radial treemap, ring bracket

REPRESENTATION DESCRIPTION

A sunburst chart displays hierarchical and part-to-whole relationships across multiple tiers of categorical dimensions. In contrast to the dendrogram, the sunburst uses layers of concentric rings, one layer for each tier. Starting from the centre ‘parent’ tier, the outward adjacency of the constituent parts of each tier represents the ‘parent-and-child’ hierarchical composition. Each ring layer is divided into proportional quantitative parts for each constituent category across that tier. The size of the quantitative parts is represented by the size of a circular arc section (in length; width is constant). Colours are often used to achieve further categorical distinction.

EXAMPLE
Showing a breakdown of the types of companies responsible for extracting different volumes of carbon-based fuels through various activities.

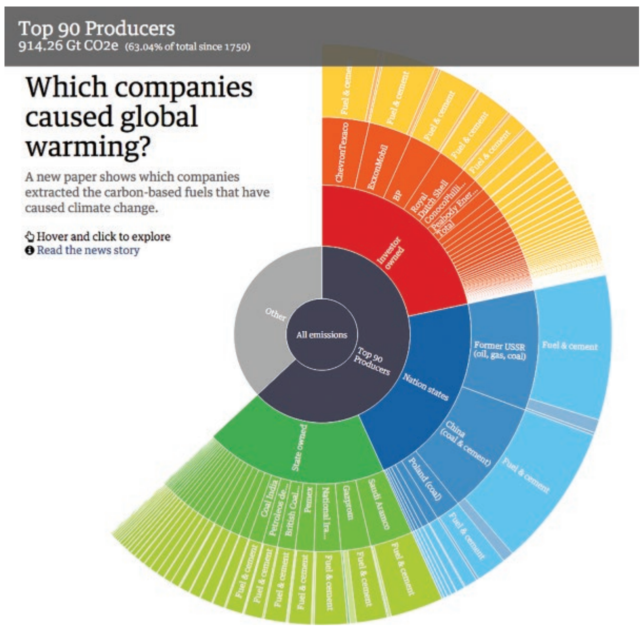


Figure 6.27 Which Fossil Fuel Companies are Most Responsible for Climate Change?, by Duncan Clark and Robin Houston (Kiln) published in the *Guardian*, drawing on work by Mike Bostock and Jason Davies

PRESENTATION TIPS

INTERACTIVITY: Interactive features that enable selection events to trigger annotated tooltips can be useful, providing direct value labels and details. There may also be scope for modifications to temporal dimensions, changing the sizes and colouring accordingly, or zooming techniques to get a closer view of small constituent parts.
ANNOTATION: Effective direct value labelling becomes difficult as the constituent parts get smaller, so only the most prominent values may be annotated, especially if interactivity is not available. Any colours used must be explained through the inclusion of a legend.
COMPOSITION: Sometimes, the hierarchical tiers do not necessarily have a parent–child relationship, so their ordering can be legitimately switched around. Therefore, careful decisions are needed about the most logical hierarchical sequencing given the subject matter and enquiry. There is also scope for arranging the sequencing of constituent parts within each tier in a meaningful way.

VARIATIONS & ALTERNATIVES

Whereas the sunburst chart uses a radial layout, the ‘icicle chart’ uses a vertical, linear layout starting from the top and moving downwards. The choice of a linear or radial tree structure will be informed largely by the space you have to work in, as well as by the legitimacy of the cyclical nature of the content in your data. A variation of the sunburst chart would be the ‘ring bracket’. This might show a hierarchical sequence of data related to subjects like sporting competitions showing a knock-out sequence.

DENDROGRAM

ALSO KNOWN AS Radial tree, layout tree, cluster tree, tree hierarchy

C **H** R T S

HIERARCHIES

REPRESENTATION DESCRIPTION

A dendrogram is a node–link diagram that displays hierarchical relationships across multiple tiers of categorical dimensions. It displays a hierarchy based on multi-generational 'parent-and-child' relationships. Starting from a singular origin root node (or 'parent') each subsequent set of constituent 'child' nodes, a tier below and represented by points, is connected by lines (curved or straight) to indicate the existence of a relationship. Each constituent node may then have further constituencies represented in the same way, continuing through to the lowest tier of detail.

EXAMPLE Showing a breakdown of the job skills required by workers in the UK job market organised by a hierarchical taxonomy of skill clusters.

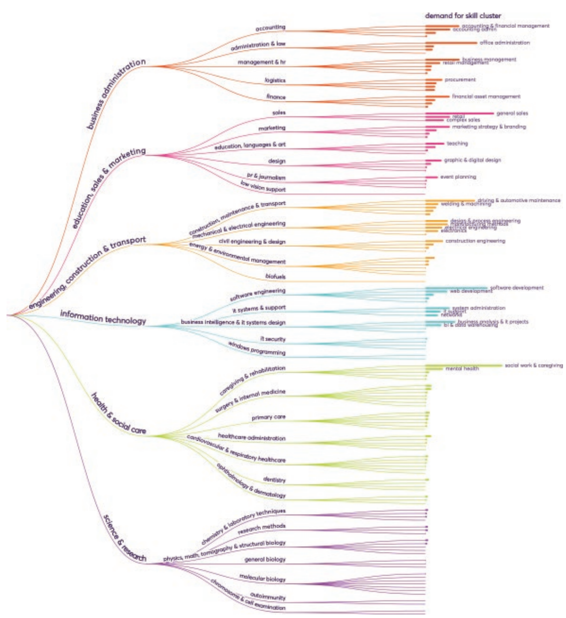


Figure 6.28 Making Sense of Skills: A UK Skills Taxonomy, by Dr Cath Sleeman

VARIATIONS & ALTERNATIVES

Variations of the dendrogram involve incorporating some additional quantitative representation such as using the length or width of connecting lines and, on replacing the point marks for each node, varying the size of node shapes. An alternative approach would be to consider the 'sunburst chart' which would show a part-to-whole relationship across the constituent categories in each hierarchical tier.

PRESENTATION TIPS

INTERACTIVITY: A useful interactive feature would be to enable filtering or highlighting of branches of interest and selection options for revealing tooltips if labelling is too difficult to accommodate elegantly.

ANNOTATION: If labelling is required, depending on the number of tiers and nodes, the size of the text will need to be carefully considered to ensure readability and minimise the effect of clutter.

COLOUR: Colour would be an optional attribute for accentuating certain nodes or applying further detail of categorisation. The colour of the connecting lines is usually based on a neutral option like black or grey.

COMPOSITION: The layout can be based on either a linear tree (typically left to right, top to bottom) or radial tree (outwards from the centre) structure. Sometimes, the hierarchical tiers do not necessarily have a parent–child relationship, so their ordering can be legitimately switched around. Therefore, careful decisions are needed about the most logical hierarchical sequencing given the subject matter and enquiry. There is also scope for arranging the sequencing of constituent parts within each tier in a meaningful way.

VENN DIAGRAM

ALSO KNOWN AS Set diagram, Euler diagram (wrongly)

C H R T S

HIERARCHIES

REPRESENTATION DESCRIPTION

A Venn diagram shows collections of and relationships between multiple sets. It typically uses circular containers to represent all independent and intersecting permutations with value labels or point marks used to place all category items in the appropriate container. The size of the contained area is (typically) not important; what is important is in which containing region an item resides. Variations in the attributes of colour or symbol may be commonly used to represent further unique distinction among the items displayed.

EXAMPLE

Comparing sets of permutations for legalities around marijuana usage and same-sex marriage across states of the USA as at 2014.

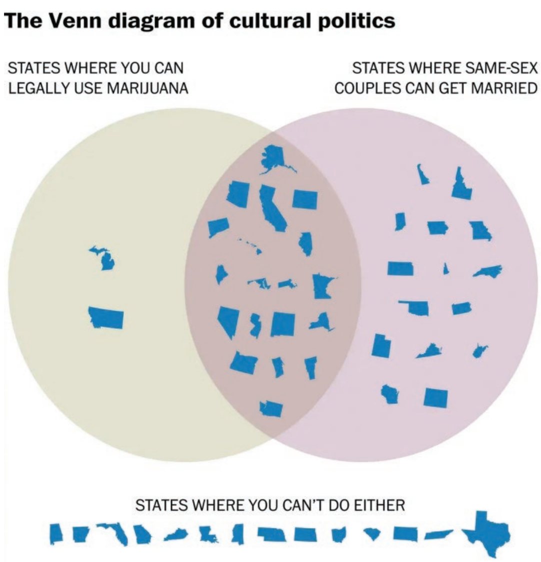


Figure 6.29 This Venn Diagram Shows Where You Can Both Smoke Weed and Get a Same-sex Marriage, by Phillip Bump (*Washington Post*)

PRESENTATION TIPS

ANNOTATION: The main annotation feature required will be to make clear which containers relate to which set or membership grouping. When the permutations increase in number (e.g. three- or four-way Venns) it can be hard to accommodate reasonable labels in each possible container.

COLOUR: Colour is often used to create more immediate distinction between the intersections and independent parts or members of each container, especially when multi-way Venns are being used.

COMPOSITION: As the attributes of size and shape of the containers are of no significance there is more flexibility to manipulate the display to modify the layout to accommodate the number or size of items in each container group. While it is theoretically possible to exceed five-way Venn diagrams, the ability of readers to make sense of such displays diminishes significantly.

VARIATIONS & ALTERNATIVES

A common variation or alternative to the Venn (but often mistakenly called a Venn) is the 'Euler diagram'. The difference is that an Euler diagram does not need to present every possible intersection and independency from all categorical sets. A different approach to visualising sets (especially larger numbers) can be achieved using the 'UpSet' technique, which uses a matrix layout to present all possible set combinations and then a second, aligned method like a bar chart to reveal a quantitative count for each set.



SCATTER PLOT

ALSO KNOWN AS Scatter graph, scatter chart

REPRESENTATION DESCRIPTION

A scatter plot displays the relationship between two quantitative variables for different category items. The display is formed by point marks for each item, plotted positionally along each quantitative axis. Sometimes attributes of colour hue are used to distinguish categorical dimensions across all items.

EXAMPLE Exploring the relationship between life expectancy and the percentage of healthy years across all countries.

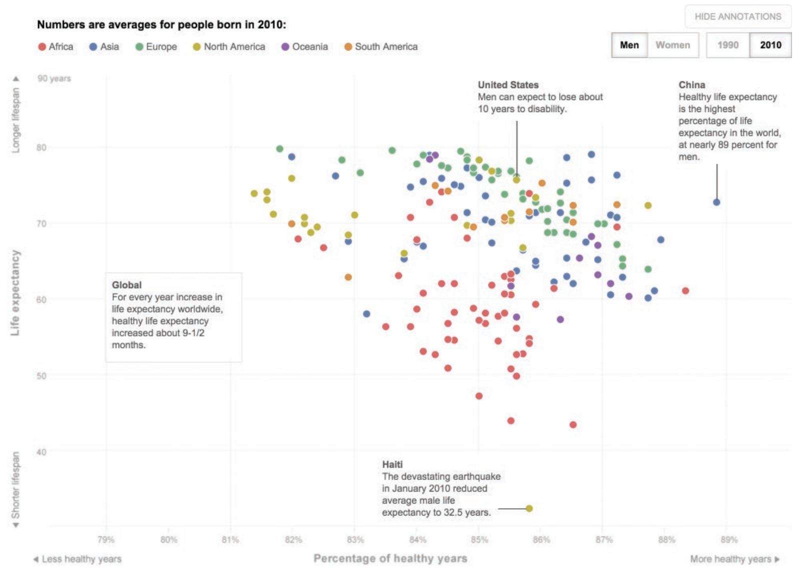


Figure 6.30 How Long Will We Live – And How Well?, by Bonnie Berkowitz, Emily Chow and Todd Lindeman (*Washington Post*)

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. Reference lines, such as a trend line of best fit, might also aid interpretation. If you include axis-scale labels you should not need to label each value directly, as this will lead to label overload. Direct labelling will normally be restricted to noteworthy points only. Any colours used must be explained through the inclusion of a legend.

COLOUR: To overcome occlusion caused by plotting several marks at the same value position, you could use unfilled or semi-transparent filled circles to help convey value frequency.

COMPOSITION: As the representation of the quantitative values is encoded through position along a scale, the quantitative axis does not need to have a zero origin, unless this is meaningful to the subject. If you do not commence from an origin of zero, this will need to be clearly annotated. Ideally a scatter plot will have a squared aspect ratio (equally tall as it is wide) to help patterns surface more evidently. If one quantitative variable (e.g. weight) is likely to be affected by the other variable (e.g. height), it is general practice to place the former on the y-axis and the latter on the x-axis. If you have to use a logarithmic quantitative scale on either or both axes, you need to make this clear to viewers.

VARIATIONS & ALTERNATIVES

A 'ternary plot' is a variation of the scatter plot through the inclusion of a third quantitative variable axis. The 'bubble plot' also incorporates a third quantitative variable, this time through encoding the size of a geometric shape (replacing the point marker). A 'scatter plot matrix' involves a single view of multiple scatter plots presenting different combinations of plotted quantitative variables, used to explore possible relationships among larger multivariate datasets. A 'connected scatter plot' compares the shifting state of two quantitative measures over time.

BUBBLE PLOT

ALSO KNOWN AS Bubble chart

C H **R** T S

CORRELATIONS

REPRESENTATION DESCRIPTION

A bubble plot displays the relationship between three quantitative variables for different category items. In contrast to the scatter plot, the bubble plot uses shape marks (usually circles) for each category item, plotted positionally along each quantitative axis with the variation in size of each mark representing a third quantitative measure. Sometimes attributes of colour hue are used to distinguish categorical dimensions across all items.

EXAMPLE

Exploring the relationship between ecological footprint and biocapacity by country.

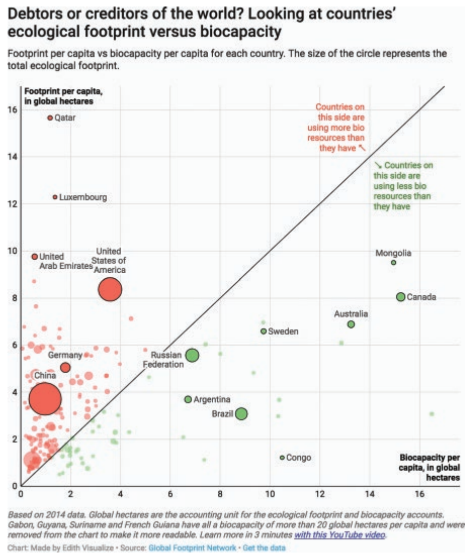


Figure 6.31 Debtors or Creditors of the World? Looking at Countries' Ecological Footprint Versus Biocapacity, by Lisa Rost and Edith Maulandi

VARIATIONS & ALTERNATIVES

If the third quantitative variable is removed, the chart type would revert to a 'scatter plot'. Variations on the bubble plot might see the use of different geometric shapes as the markers, maybe introducing extra meaning through the shape and dimensions used.

PRESENTATION TIPS

INTERACTIVITY: A useful interactive feature would be to enable filtering or highlighting of certain categorical items, especially if there are several distinct categories and lots of items to make sense of. Furthermore, selection options for revealing tooltips can be helpful if direct labelling is too difficult to accommodate elegantly.

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. Reference lines, such as a trend line of best fit, might also aid interpretation. If you include axis-scale labels you should not need to label each value directly, as this will lead to label overload. Direct labelling will normally be restricted to noteworthy points only. Any colours used must be explained through the inclusion of a legend.

COLOUR: If colours are being used to distinguish the different categories, ensure these are as visibly different as possible. When data values are especially diverse in range, the size of shapes may vary from very small to quite large. The largest shapes may overlap, in spatial terms, with other values or even hide them completely. The use of outline borders and semi-transparent colours can help avoid the effect of total occlusion.

COMPOSITION: As the representation of the quantitative values is encoded through position along a scale, the quantitative axis does not need to have a zero origin, unless this is meaningful to the subject. If you do not commence from an origin of zero, this will need to be clearly annotated. Ideally a bubble plot will have a squared aspect ratio (equally tall as it is wide) to help patterns surface more evidently. If one quantitative variable (e.g. weight) is likely to be affected by the other variable (e.g. height), it is general practice to place the former on the y-axis and the latter on the x-axis. If you have to use a logarithmic quantitative scale on either or both axes, you need to make this clear to viewers. The geometric accuracy of the shape mark size calculation is paramount: it is the area you are modifying, not the diameter/radius.

NETWORK DIAGRAM

ALSO KNOWN AS Node–link diagram, graph, hairball graph, social network

C H **R** T S

CONNECTIONS

REPRESENTATION DESCRIPTION

Node–link diagrams display relationships through the connections between data items. The common version of this type of diagram displays items as nodes, represented by point marks, with links or edges (represented by lines) depicting the existence of a connection. These connecting lines will sometimes encode an attribute of direction to indicate the influencer relationship. Replacing point marks with a geometric shape and using attributes of size is a further variation. In some versions a further quantitative weighting is applied to show the relationship strength, maybe through increased line width. Attributes of colour may be used to indicate a quantitative value (e.g. number of connections) and/or some categorical classification.

EXAMPLE Exploring the connections of voting patterns for Democrats and Republicans across all members of the US House of Representatives from 1949 to 2012.

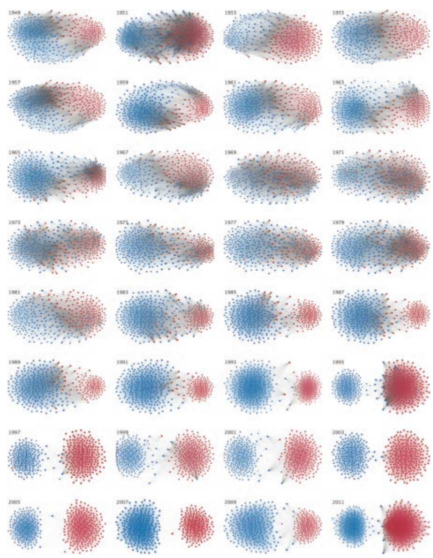


Figure 6.32 The Rise of Partisanship and Super-cooperators in the U.S. House of Representatives, visualisation by Mauro Martino, authored by Clio Andris, David Lee, Marcus J. Hamilton, Mauro Martino, Christian E. Gunning, and John Armistead Selde

VARIATIONS & ALTERNATIVES

There are many derivatives of the node–link diagram, as explained, based on variations in the use of different attributes. 'Hive plots' and 'BioFabric' offer alternative approaches based on replacing nodes with vertices.

PRESENTATION TIPS

INTERACTIVITY: Node–link diagrams are particularly useful when offered with interactive features, enabling the user to interrogate and manipulate the display to facilitate visual exploration. The option to apply filters to reduce the busyness of the display, and maybe enable isolation of individual node connections, can help viewers to focus on specific parts of the network rather than face the whole system at once.

ANNOTATION: The complexity revealed by these diagrams is often a reflection of the underlying subject complexity, so it can be helpful to use annotation to surface key observations about significant clusters or label those items with the most connections.

COLOUR: Aside from the possible categorical colouring of each node, decisions need to be made about the colour of the connecting lines, especially with regard to how prominent these links will be in contrast to the nodes.

COMPOSITION: Composition decisions will be so varied for any network diagram depending on the complexity and volume of the data and the output constraints around space and consumption format. There are several common algorithmic treatments used to compute custom arrangements to optimise network displays, such as force-directed layouts using the physics of repulsion and springs to amplify relationships. There are also simplifying techniques such as edge bundling to aggregate or summarise multiple similar links.

SANKEY DIAGRAM

ALSO KNOWN AS Alluvial diagram

REPRESENTATION DESCRIPTION

Sankey diagrams display categorical composition and flows of quantitative relationships between different major ordinal dimensions. The original application of the Sankey diagram displayed flow relationships over many discrete ordinal stages, but it would be reasonable to say most common forms involve a two-sided parallel display. The two sides represent different states of a paired, ordinal relationship, such as input vs output, or time A vs time B. On each side there is effectively a stacked bar chart displaying proportionally sized and differently coloured (or spaced apart) constituent parts of each whole. These might show categorical breakdowns of income vs categories of expenditure or the categorical composition of some whole in a before and after state. Curved bands join each side of the display to represent the connecting categories (origin and destination) with proportionally sized thickness representing the quantitative flow of this relationship.

EXAMPLE
Showing a breakdown of reasons for animals being brought into a shelter and a breakdown of the related outcomes of each animal after one month.

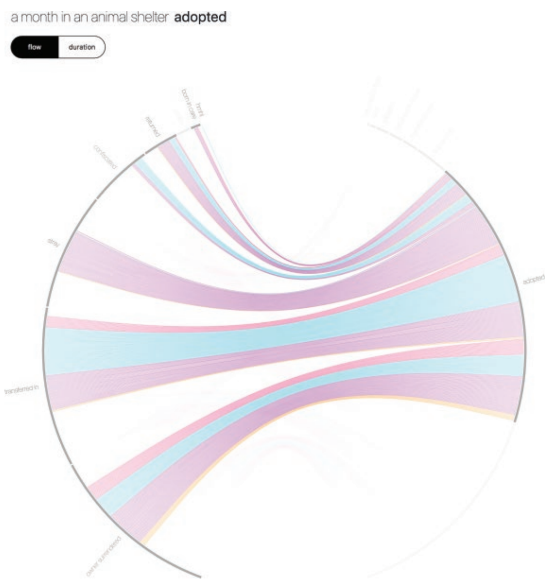


Figure 6.33 A Month in an Animal Shelter, by Sarah Campbell

VARIATIONS & ALTERNATIVES

The Sankey is closely related to the 'alluvial diagram', which tends to show changes in composition and flow over time and often across multiple stages (rather than just the common paired structure of the Sankey). These days, the labels are often applied interchangeably. A 'chord diagram' is a variation that uses a radial display to enable certain origins and destinations to be one and the same. Showing how component parts have changed over time could just be displayed using a 'stacked area chart'. Plotting composition and flow can also be applied to a spatial display to create a variation of the geographical 'flow map'.

PRESENTATION TIPS

INTERACTIVITY: Sankey diagrams are particularly useful when offered with interactive features, enabling the user to interrogate and manipulate the display to facilitate visual exploration. The option to apply filters to reduce the chaos of the visual and enable isolation of individual or groups of flows helps users to focus on specific relationships of interest. Interactively enabled labelling can also be beneficial as direct labelling is difficult to incorporate elegantly.

ANNOTATION: Direct labelling will normally be restricted to noteworthy points only. Any colours used must be explained through the inclusion of a legend.

COLOUR: Colouring is often used visually to indicate the categories of the connecting bands, though this can become complicated by an origin categorical colour joining to a different destination colour. A sense of this change can be conveyed by blending the origin and destination colours half-way across.

COMPOSITION: The main arrangement decisions concern sorting. Firstly, by deriving as much logical meaning from the categorical values within each stack and, secondly, by deciding on the sorting of the connecting lines in the z-dimension – if many lines are crossing, there is a need to think about which will be on top and which will be below. There is no significant difference between a landscape or portrait layout; it will depend on the subject-matter 'fit' and the space within which you have to work. Sometimes the stacks on each side are curved and appear more like stacked arcs.

CHORD DIAGRAM

ALSO KNOWN AS Radial Sankey diagram, radial network, arc diagram

REPRESENTATION DESCRIPTION

A chord diagram displays relationships through connections between and within category items. The diagram is formed around a radial display with categories located around the edge: either shown as individual nodes or as arc segments proportionally sized around the circumference to represent a part-to-whole breakdown. Emerging inwards from each origin position are curved lines that join with related categorical destinations around the edge. The connecting lines are proportionally sized according to a quantitative measure and a directional or influencing relationship is often indicated. Attributes of colour hue are commonly used to distinguish different category groupings visually.

EXAMPLE Exploring the connections of migration between and within ten world regions based on estimates across five-year intervals between 1990 and 2010.



Figure 6.34 The Global Flow of People, by Nikola Sander, Guy J. Abel and Ramon Bauer

PRESENTATION TIPS

INTERACTIVITY: Chord diagrams are particularly useful when offered with interactive features, enabling the user to interrogate and manipulate the display to facilitate visual exploration. The option to apply filters to reduce the chaos of the visual and enable isolation of individual or groups of flows helps users to focus on specific relationships of interest. Interactively enabled labelling can also be beneficial as direct labelling is difficulty to incorporate elegantly.

ANNOTATION: Direct labelling will normally be restricted to noteworthy points only. Any colours used must be explained through the inclusion of a legend.

COLOUR: Aside from the categorical colouring of each node, decisions need to be made about the colour of the connecting lines, especially on deciding how prominent these links will be in contrast to the nodes. Sometimes the connections will match the origin or destination colours, or they will combine the two (with a start and end colour blend to match the relationship).

COMPOSITION: The main arrangement decisions concern sorting. Firstly, by deriving as much logical meaning from the categorical values within each stack and, secondly, by deciding on the sorting of the connecting lines in the z-dimension – if many lines are crossing, there is a need to think about which will be on top and which will be below. Showing the direction of connections can be achieved using arrowheads or colour lightness variation. One common, subtle solution is to pull the destination join away from the edge of the destination arc to contrast with connecting lines that emerge directly from an origin.

VARIATIONS & ALTERNATIVES

Variations of the chord diagram would be to consider using a single baseline axis, placing all category items along it and forming connections between using semi-circular arcs. Additionally, a 'Sankey diagram' would be relevant if there are distinct origins and destination relationships to reveal.

LINE CHART

ALSO KNOWN AS Fever chart, stock chart

C H R **T** S

TRENDS

REPRESENTATION DESCRIPTION

A line chart shows how quantitative values have changed over time for different categorical items. Line charts are typically structured around a continuous temporal x-axis and quantitative y-axis with values plotted using point marks at relevant coordinates. Connecting lines join up adjacent and related categorical items to form slopes which are then extended along the full timescale to display a complete change over time. Multiple categories can be displayed in the same view, each represented by a discrete line often with categorical or editorial colour attributes. The connecting lines are typically straight, but sometimes curved line 'interpolation' may be applied to help emphasise a general trend above precise point reading.

EXAMPLE Showing cumulative runs scored in Test matches by English Test batsmen between 1947 and 2018.

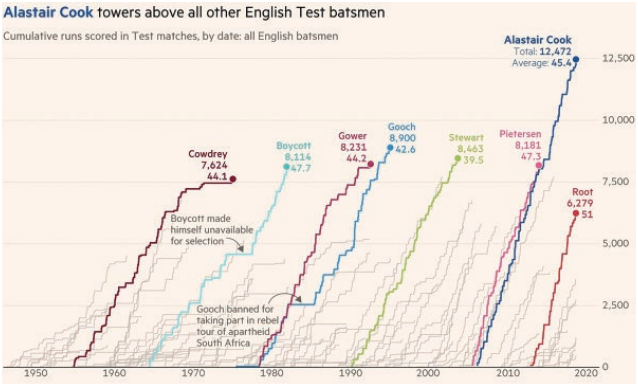


Figure 6.35 Cricketer Alastair Cook Plays His 161st and Final Test Match, by *Financial Times* / John Burn-Murdoch

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be especially helpful if you have many discrete categorical lines and wish to enable the user to isolate a certain category of interest, either through filtering to exclude the others or using a contrasting colour to emphasise its shape among the rest.

ANNOTATION: Sometimes the point mark is quite pronounced, to aid value judgements and possibly to provide space for a value label, but on most occasions only the position of the connecting lines is displayed. Ranking labels can be derived from the vertical position along the scale so direct labelling is usually unnecessary. The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. You might choose to annotate specific values of interest (highest, lowest, specific milestones). Any colours used must be explained through the inclusion of a legend. If the shape of the data presents an opportunity, you might consider directly labelling each or specific category lines, at the first or last point mark position.

COLOUR: When many categories are shown, rather than colouring each line with a distinct categorical classification, it may only be viable to emphasise lines of interest using colour hue or saturation.

COMPOSITION: The chart's dimensions will need to be carefully considered, specifically the aspect ratio formed by its height and width. The upward and downward slopes can seem more significant if the chart width is narrow and less significant if it is more stretched out. There is no single practical rule to obey other than using common sense to ensure you do not overly amplify or underplay features of your data. The sequencing of values tends to follow a chronological left-to-right direction for the time-based x-axis and low values rising up to high values on the y-axis; you will need a good (and clearly annotated) reason to break this convention. Line charts do not always require the quantitative axis origin to start from zero, as the size of a value is represented through position along a scale rather than the size of a line or shape. If zero has significance for the interpretation of the trends portrayed, given the subject matter, then you should start the baseline at zero.

VARIATIONS & ALTERNATIVES

Variations of the line chart may include the 'bump chart', to show rankings over time, and the 'slope graph', to compare trends over two points in time. 'Spark lines' are mini line charts that aim to occupy only a word's length amount of space. They are often seen in dashboards where space is at a premium and there is a desire to optimise the density of the display. An alternative would to use the 'bar chart' when you have quantities for discrete periods (such as totals over a monthly period) rather than a purely continuous series of point-in-time measurements.

BUMP CHART

ALSO KNOWN AS Bumps chart, rank chart

REPRESENTATION DESCRIPTION

A bump chart shows how quantitative values have changed over time for different categorical items, where the quantitative values are ranking measurements. These charts are typically structured around a continuous temporal x-axis and quantitative y-axis with values plotted using point marks at relevant coordinates. Connecting lines join up adjacent and related items to form slopes which are then extended along the full timescale to display a complete change over time. A common extension of the bump charts uses variation in the size (width) of each line to represent a quantitative measure, usually the absolute value that informs the ranking measurement. Multiple categories are commonly displayed in the same view, each represented by a discrete line often with categorical or editorial colour attributes. The connecting lines are typically straight, but sometimes curved line 'interpolation' may be applied to help emphasise a general trend above precise point reading.

EXAMPLE Showing changes in rank of the most politically important topics for Germans between 1998 and 2017.

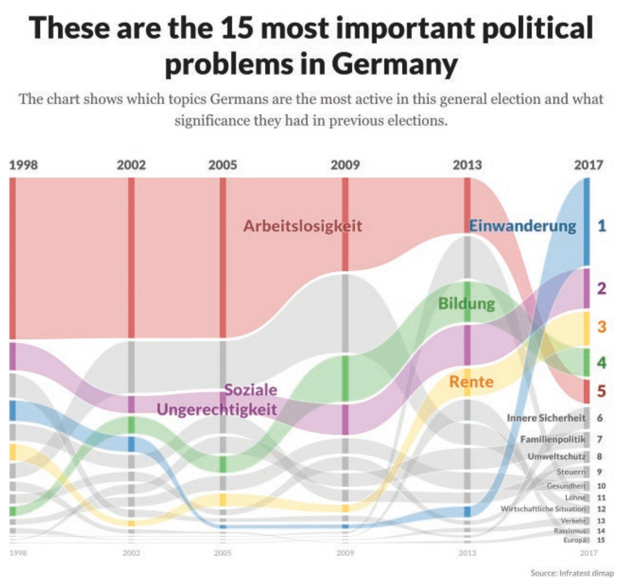


Figure 6.36 These are the 15 Most Important Political Problems in Germany, [Translated] by *Berliner Morgenpost*

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be especially helpful if you have many discrete categorical lines and wish to enable the user to isolate a certain category of interest, either through filtering to exclude the others or using a contrasting colour to emphasise its shape among the rest.

ANNOTATION: Sometimes the point mark is quite pronounced, to aid value judgements and possibly to provide space for a value label, but on most occasions only the position of the connecting lines is displayed. Ranking labels can be derived from the vertical position along the scale so direct labelling is usually unnecessary. You might choose to annotate specific values of interest (highest, lowest, specific milestones). Any colours used must be explained through the inclusion of a legend. If the shape of the data presents an opportunity, you might consider directly labelling each or specific category lines, at the first or last point mark position, or even both.

COLOUR: When many categories are shown, rather than colouring each line with a distinct categorical classification, it may only be viable to emphasise lines of interest using colour hue or saturation.

COMPOSITION: The sequencing of values tends to follow a chronological left-to-right direction for the time-based x-axis and highest ranking values dropping to lowest ranking values on the y-axis; you will need a good (and clearly annotated) reason to break this convention.

VARIATIONS & ALTERNATIVES

Consider alternatives like 'line charts' and 'area charts' if the ranking measurement is of secondary interest to plotting absolute quantitative values.

C H R T S

TRENDS

SLOPE GRAPH

ALSO KNOWN AS Slope chart, parallel coordinates

REPRESENTATION DESCRIPTION

A slope graph shows how quantitative values have changed over two points in time for different category items. The display is based on (typically) two parallel quantitative axes with a common value range. A line is plotted for each category connecting the two axes together with the vertical position on each axis representing the respective quantitative values. These connecting lines form slopes that indicate the upward, downward or stable trend between the two temporal axes. Attributes of colours are often used to distinguish different categorical lines or to surface major trends among the items plotted.

EXAMPLE

Showing changes in the share of power sources across all US states between 2004 and 2014.

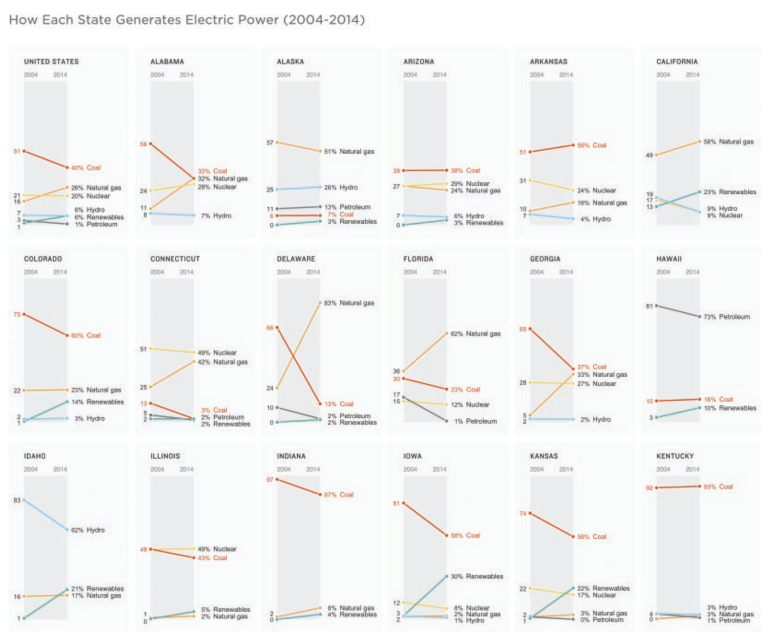


Figure 6.37 Coal, Gas, Nuclear, Hydro? How Your State Generates Power

Source: U.S. Energy Information Administration, Credit: Christopher Groskopf, Alyson Hurt and Avie Schneider (NPR)

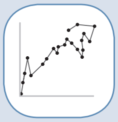
PRESENTATION TIPS

INTERACTIVITY: Depending on the number of category values being presented, slope graphs can become quite busy, especially if there are bunches of similar quantitative values with slope transitions. This also causes a problem with accommodating multiple labels on the same value. On these occasions you might find interactive features useful to enable filtering of certain items, to exclude others or to highlight a selection. Discovering value labels of each item through interactive tooltips can also be beneficial.

ANNOTATION: Labelling of each category item on both sides will often be necessary, though this can be challenging composition-wise when there are several items positioned in close proximity. You might therefore choose to annotate only specific values of interest (highest, lowest, of editorial interest). The parallel axes will need clear labels to explain the respective points in time. Any colours used must be explained through the inclusion of a legend.

VARIATIONS & ALTERNATIVES

Rather than comparing two points in time, some variations in the application of a slope graph are used to show the relationship between discrete quantitative variables for related category items. In this case the connecting line is not indicative of a trend, rather a join to connected related items. This approach can also lead to the slope graph being extended beyond just two parallel axes and thus evolving into the technique known as ‘parallel coordinates’. An alternative chart type would be to consider the ‘connected dot plot’ which can also show comparisons of quantities for two points in time across multiple category items.



CONNECTED SCATTER PLOT

ALSO KNOWN AS Trail chart, comet chart

C H R **T** S

TRENDS

REPRESENTATION DESCRIPTION

A connected scatter plot displays the relationship between two quantitative measures over time. The display is formed of two quantitative x- and y-axes and with the values represented by point marks at the respective coordinates, one for each measurement over time. The individual points are then connected (think of a dot-to-dot drawing puzzle) using lines joining each consecutive point in time to form a sequence of change.

EXAMPLE Showing changes in the daily price and availability of Super Bowl tickets on the secondary market four weeks prior to the event across five Super Bowl finals.



Figure 6.38 Holdouts Find Cheapest Super Bowl Tickets Late in the Game, by Alex Tribou, David Ingold and Jeremy Diamond (Bloomberg Visual Data)

VARIATIONS & ALTERNATIVES

The 'comet chart' is to the connected scatter plot what the 'slope graph' is to the 'line chart' – a summarised view of the relationship between two quantitative measures over two points in time. The comet aspect is demonstrated through the cone shape of the connecting line, with the more recent period of time generally having a thicker width. A variation of the connected scatter plot is simply the 'scatter plot' where there is no time dimension or elements of connectedness between points.

PRESENTATION TIPS

INTERACTIVITY: The biggest challenge is making the connections and the sequence as visible as possible. This becomes much harder when values change very little and/or they loop back to previous positions, crossing back over themselves. It is especially hard to label the sequential time values elegantly. One option to overcome this is through animated sequences which might build up the display, connecting one line at a time and unveiling the date labels as time progresses. It is often the case that only one series will be plotted. However, interactive options may allow the user to overlay one or more for comparison, switching them on and off as required.

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. If you can elegantly include direct labels to each point value, indicating the time period it relates to, then this can be helpful. Connected scatter plots are unfamiliar to many audiences and it can be quite demanding to learn how to read them. It may be necessary to provide a 'how to read' guide illustrating what the axis values represent and what it means when connecting lines are moving in different directions. Also consider labelling parts of the chart region that carry particular meaning, so if a connecting line moves into that region, the interpretation of what this means can be accelerated.

COLOUR: Colour might be used to explain the temporal status of the connecting lines, for instance using a faded colour for the past and a more vivid colour for the present. Otherwise, you might use attributes of colour to accentuate certain sections of a sequence that might warrant particular attention.

COMPOSITION: As the representation of the quantitative values is encoded through position along a scale, the quantitative axis does not need to have a zero origin, unless this is meaningful to the subject. If you do not commence from an origin of zero, this will need to be clearly annotated. Ideally a scatter plot will have a squared aspect ratio (equally tall as it is wide) to help patterns surface more evidently. If one quantitative variable (e.g. weight) is likely to be affected by the other variable (e.g. height), it is general practice to place the former on the y-axis and the latter on the x-axis.

AREA CHART

ALSO KNOWN AS Density plot

C H R T S

TRENDS

REPRESENTATION DESCRIPTION

An area chart shows how quantitative values have changed over time for a single categorical item. The charts are typically structured around a continuous temporal x-axis and quantitative y-axis with values plotted using point marks at relevant coordinates. Connecting lines join up adjacent and related items to form slopes which are then extended along the full timescale to display a complete change over time. The connecting lines are typically straight, but sometimes curved line 'interpolation' may be applied to help emphasise a general trend above precise point reading. To accentuate the shape of the trends, the area beneath the line is filled with colour, which means the height of the area at any given point also reveals its quantity.

EXAMPLE
Showing changes in the average weekly price ($ per barrel) of Brent crude oil between 2008 and 2018.

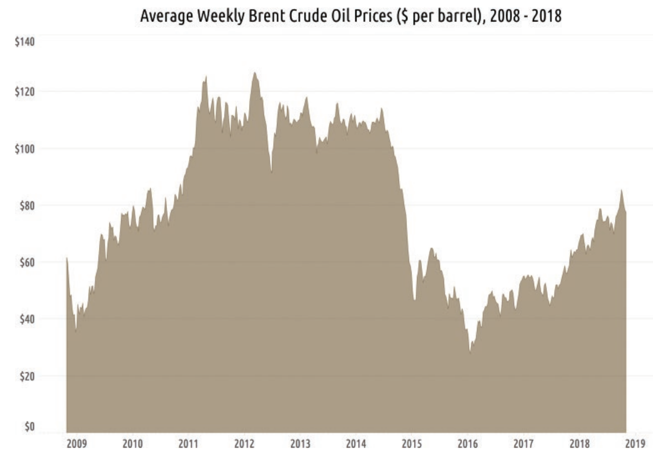


Figure 6.39 Crude Oil Prices for Brent (Dollars per Barrel) 2008–2018

PRESENTATION TIPS

ANNOTATION: Sometimes the point mark is quite pronounced, to aid value judgements and possibly to provide space for a value label, but on most occasions only the position of the connecting lines is displayed. The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. You might choose to annotate specific values of interest (highest, lowest, specific milestones).

COMPOSITION: The chart's dimensions will need to be carefully considered, specifically the aspect ratio formed by its height and width. The upward and downward slopes can seem more significant if the chart width is narrow and less significant if it is more stretched out. There is no single practical rule to obey other than using common sense to ensure you do not overly amplify or underplay features of your data. The sequencing of values tends to follow a chronological left-to-right direction for the time-based x-axis and low values rising up to high values on the y-axis; you will need a good (and clearly annotated) reason to break this convention. Unlike the line chart, the quantitative axis for area charts must have an origin of zero as it is the height of the coloured area under the trend line that is used to perceive the quantitative values.

VARIATIONS & ALTERNATIVES

A variation of the area chart is the 'stacked area chart', which can be used to show how multiple categories form a whole and how this composition changes over time. The stacks may amount to an absolute total or form a 100% proportion view. A 'density plot' appears like an area chart but is used to show the distribution of values across a quantitative axis, rather than a time axis. Another variation is the 'horizon chart', which is based on an area chart but for space-limited contexts. Values that exceed an imposed fixed maximum y-axis range are coloured to indicate different bands of magnitudes, with the extremes usually darker. Like slicing layers off a mountain, each distinct band of values above the maximum y-axis range is chopped off and dropped down to the baseline in front of its foundation base. The final effect shows overlapping layers of increasingly darker colour-shaded areas occupying the same vertical space. An alternative may be simply to consider the 'line chart', especially if you want to compare against several discrete categorical items.

STACKED AREA CHART

ALSO KNOWN AS Area chart, horizon chart

REPRESENTATION DESCRIPTION

A stacked area chart shows how quantitative values have changed over time for multiple categorical items. These charts are typically structured around a continuous temporal x-axis and quantitative y-axis with values plotted using point marks at relevant coordinates. Connecting lines join up adjacent and related items to form slopes which are then extended along the full timescale to display a complete change over time. The connecting lines are typically straight, but sometimes curved line 'interpolation' may be applied to help emphasise a general trend above precise point reading. To accentuate the shape of the trends, the area beneath the line is filled with colour, which means the height of the area at any given point also reveals its quantity. When there are multiple discrete categories, separate stacked areas, sized in height proportionally to their shifting values, are distinguished through distinct stacked regions often coloured to establish their categorical association. The resulting display reveals how a part-to-whole relationship changes over time.

Example
Showing the average trends of bike share usage across the bike share stations of Philadelphia.

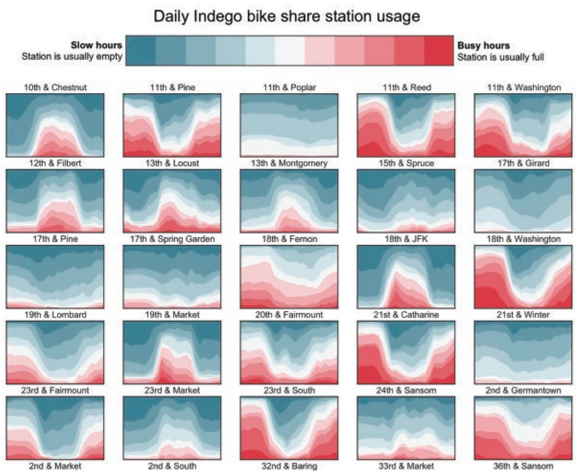


Figure 6.40 Daily Indego Bike Share Station Usage, by Randy Olson (@randal_olson)(www.randalolson.com)

VARIATIONS & ALTERNATIVES

The main variation in stacked area charts will be based on the quantitative values plotted and whether they are representative of an absolute total or a proportional total forming a 100% whole. Rather than stacking categories you might consider using small multiples of single-category area charts, especially as this will display each from a common baseline and therefore make judgements of shape and size a little easier. An alternative may be simply to consider the 'line chart' formed of multiple lines for discrete categorical items.

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be especially helpful if you have many discrete categorical stacks and wish to enable the user to isolate a certain category of interest, either through filtering to exclude the others or using a contrasting colour to emphasise its shape among the rest. Revealing the quantitative value, time and category label at any point on the chart through a selectable tooltip can also be useful.

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the quantitative values. You might choose to annotate specific values of interest (highest, lowest, specific milestones). Directly labelling the discrete category stacks can be helpful, otherwise use a clear colour legend to explain associations.

COMPOSITION: The chart's dimensions will need to be carefully considered, specifically the aspect ratio formed by its height and width. The upward and downward slopes can seem more significant if the chart width is narrow and less significant if it is more stretched out. There is no single practical rule to obey other than using common sense to ensure you do not overly amplify or underplay features of your data. The sequencing of values tends to follow a chronological left-to-right direction for the time-based x-axis and low values rising up to high values on the y-axis; you will need a good (and clearly annotated) reason to break this convention. Unlike the line chart, the quantitative axis for stacked area charts must have an origin of zero as it is the height of the coloured areas used to perceive the quantitative values. Try to make the sorting of the categorical stacks as meaningful as possible, perhaps placing the most important values on the bottom stack to give it a consistent baseline.

STREAM GRAPH

ALSO KNOWN AS Theme river

C H R **T** S

TRENDS

REPRESENTATION DESCRIPTION

A stream graph shows how quantitative values have changed over time for multiple categorical items. The graphs are generally used when you have many concurrent, constituent categories at any given point in time and these categories may start and stop at different points in time rather than continue throughout the presented time frame. As befitting the name, the appearance of the graphs is characterised by a flowing, organic display of meandering layers. They are structured around a temporal x-axis with quantitative values plotted to quantify height above a local baseline, which is not a stable zero baseline but rather a shifting shape formed out of other category layers. Connecting lines join up adjacent and related items to form slopes which are then extended along the relevant time frame to create a unique categorical layer. The area occupied by this layer is filled with an attribute of colour to represent a further quantitative value scale or to associate with categorical classifications. The stacking arrangement of the multiple categorical layers can shift up and down the implied y-axis dimension, in order to optimise the layout, but not to indicate any notion of positive or negative values.

EXAMPLE

Showing changes in the total domestic gross takings ($US) and the longevity of all movies released between 1986 and 2008.

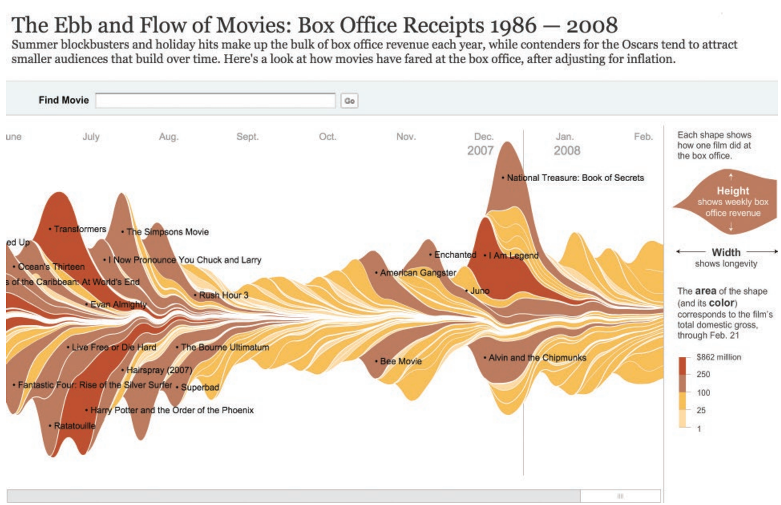


Figure 6.41 The Ebb and Flow of Movies: Box Office Receipts 1986–2008, by Mathew Bloch, Lee Byron, Shan Carter and Amanda Cox (*New York Times*)

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be especially helpful if you have many discrete categorical layers and wish to enable the user to isolate a certain category of interest, either through filtering to exclude the others or using a contrasting colour to emphasise its shape among the rest. Revealing the quantitative value, time and category label at any point on the chart through a selectable tooltip can also be useful.

ANNOTATION: Chart apparatus devices are generally of limited use in a stream graph with the priority being more on offering a general sense of pattern above precision of value reading. Direct labelling of the discrete categorical layers may be possible, depending on the shape of the data, otherwise use a clear colour legend to explain associations.

VARIATIONS & ALTERNATIVES

If you have relatively few discrete categorical items, you might consider using an alternative chart like the 'stacked area chart' or small multiples of individual 'area charts'. A 'stacked bar chart' would be a consideration, again if there are relatively few categories to include and the quantitative measurements are based on discrete periods (such as totals over a monthly period) rather than a purely continuous series of point-in-time measurements.

GANTT CHART

ALSO KNOWN AS Range chart, floating bar chart, Priestley timeline

REPRESENTATION DESCRIPTION

A Gantt chart displays time-based intervals for different categorical items. The charts are typically structured around a continuous temporal x-axis with a separate row allocated to each major categorical item. Intervals are formed by line marks positioned according to a starting point and sized through length according to a closing point in time. Point marks at each end of this line are sometimes included and presented with discrete symbols or attributes of colour to highlight their distinction. The line may also display an attribute of colour to relate to some categorical status.

EXAMPLE Showing the timeline of all current and former US national parks based on when they were officially established or designated.

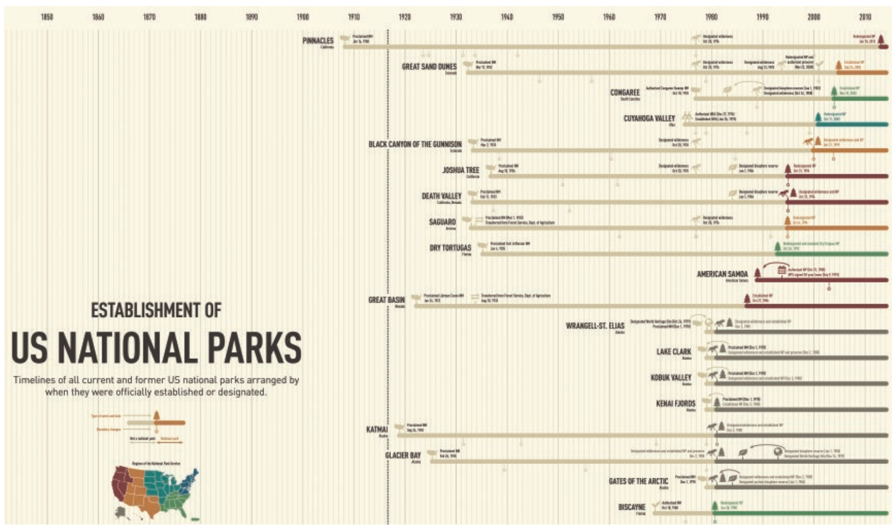


Figure 6.42 Establishment of the U.S. National Parks, by Nicholas Rougeux (www.c82net)

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the date values and durations. If you include axis-scale labels you should not need to label each bar value directly, as this will lead to label overload.

COMPOSITION: The bars should be proportionally sized according to the associated duration length – nothing more, nothing less – otherwise the perception of the bar sizes will be distorted. There is no significant difference in perception between vertically or horizontally arranged Gantt charts; it will depend on which layout makes it easier to accommodate the range of values and to read the item labels associated with each category. Landscape layouts with time chronologically sequenced from left to right would be the most common arrangement. Where possible, try to sequence the categorical items in a way that makes for the most logical reading, organised by either the start/finish dates or maybe the durations (depending on which has most relevance).

VARIATIONS & ALTERNATIVES

Gantt charts share many characteristics with the 'connected dot plot', but the measurement dimension here is of time duration rather than quantitative difference. If duration between points in time is less important than individual milestones or events, the 'instance chart' would be worth considering. Sometimes interval lines join up with other adjacent categories, rather than being bound by discrete rows. This might be representative of the merging of activities or the absence of 'discreteness' between activities, and the technique may therefore evolve more towards being a 'connected timeline'.

INSTANCE CHART

ALSO KNOWN AS Dot plot, barcode plot, strip plot

REPRESENTATION DESCRIPTION

An instance chart displays time-based events for different categorical items. It is typically structured around a continuous temporal x-axis with a separate row allocated to each major categorical item. Events are represented by point markers, plotted along the timeline, using combinations of symbols and colours to represent different types.

EXAMPLE
Showing the instances of different 'Avengers' characters appearing in Marvel's comic book titles between 1963 and 2015.

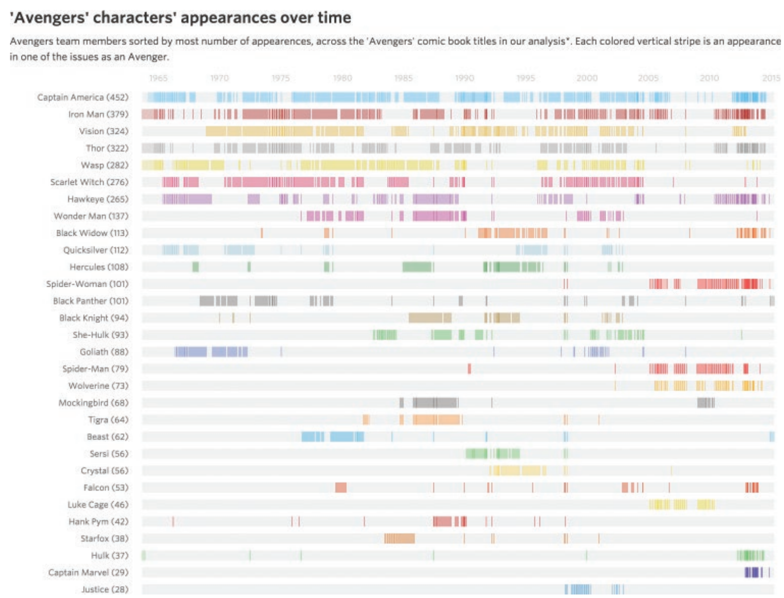


Figure 6.43 How the 'Avengers' Line-up Has Changed Over the Years, by Jon Keegan (*Wall Street Journal*)

PRESENTATION TIPS

ANNOTATION: The inclusion of chart apparatus devices like tick marks and gridlines can help increase the precision of judging the date values. If you include axis-scale labels you should not need to label each bar value directly, as this will lead to label overload. Any colours, symbols or size attributes used must be explained through the inclusion of a legend.

COMPOSITION: There is no significant difference in perception between vertically or horizontally arranged instance charts; it will depend on which layout makes it easier to accommodate the range of values and to read the item labels associated with each category. Landscape layouts with time chronologically sequenced from left to right would be the most common arrangement. Where possible, try to sequence the categorical items in a way that makes for the most logical reading, organised by either the earliest or latest points in time or maybe some measure of quantity (such as which category has the most recorded events).

VARIATIONS & ALTERNATIVES

Instance charts share many characteristics with the 'dot plot' but the measurement dimension here is of time rather than quantitative value. Variations may see the point mark replaced by a geometric shape sized to represent a quantitative measure associated with each event. If the data is more about durations and intervals between events, the 'Gantt chart' will be the best-fit option.

CHOROPLETH MAP

ALSO KNOWN AS Heat map

C H R T **S**

OVERLAYS

REPRESENTATION DESCRIPTION

A choropleth map displays quantitative values for distinct, definable spatial regions. Each region is represented by a defined polygonal shape, with each distinct shape collectively arranged to form the entire landscape. An attribute of colour is used to represent a quantitative measurement. Choropleth maps should only be used when the quantitative measure is directly associated with and continuously relevant across the spatial region on which it will be displayed. If the quantitative measure is related to the consequence of more people living in an area, consider transforming your data by standardising it as per capita or per acre (or other spatial denominator) accordingly.

EXAMPLE Mapping the results of the 2017 German general election showing the winning party for each electoral location.

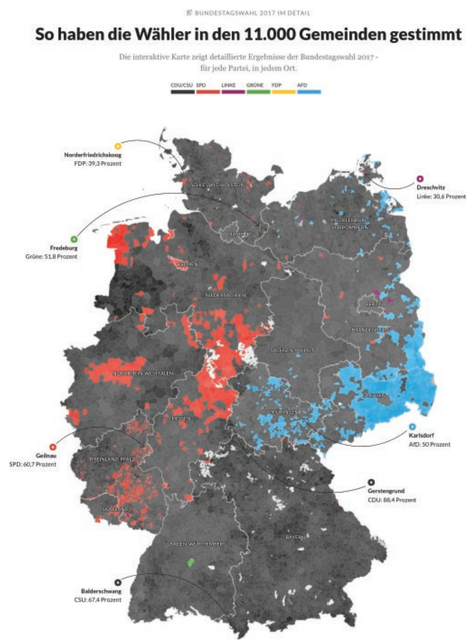


Figure 6.44 How Voters in the 11,000 Municipalities Voted, [Translated] by *Berliner Morgenpost*

VARIATIONS & ALTERNATIVES

Some choropleth maps may be used to indicate categorical association rather than quantitative measurements. Alternative thematic mapping approaches to representing quantitative values might include the 'proportional symbol map', using sized shapes over locations, and the 'dot density map', which plots a representative quantity of dots equally (but randomly) across and within a defined spatial region. 'Dasymetric mapping' is similar in approach to choropleth mapping but breaks the constituent regional areas into much smaller, more specific sub-regions better to represent the realities of the distribution of human and physical phenomena within a given spatial boundary. This might include details of individual buildings, for example.

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be especially helpful to offer selectable tooltips to view quantitative values and category or location labels for any region on the display.

ANNOTATION: Depending on the shapes of the regions displayed, direct labelling may be limited to just a number of noteworthy values. Any colours used must be explained through the inclusion of a legend. If you choose to include a detailed map image in the background, do not include any unnecessary geographic details that add no value to the spatial orientation or interpretation (e.g. roads, building structures).

COLOUR: The outline colour and stroke width for each spatial area should be distinguishable enough to define the shape but not so prominent as to dominate. Usually, a light-grey or white-coloured stroke will suffice. Sometimes variation in pattern may be included, as well as colour, to represent values that may be uncertain or incomplete. When background map images are included, consider making them semi-transparent or light in colour to avoid competition for attention with the more important data layer.

COMPOSITION: There are many different mapping projections for spatially representing the regions of the world on a plane surface. Be aware that the transformation adjustments made by some of these projections can distort the size of regions of the world, inflating their size relative to other regions, so you will need to pick a projection that is appropriate to the spatial view you are providing.

ISARITHMIC MAP

ALSO KNOWN AS Contour map, isopleth map, isochrone map

C H R T **S**

OVERLAYS

REPRESENTATION DESCRIPTION

An isarithmic map displays distinct spatial surfaces on a map that shares the same quantitative classification. The spatial definition here is not framed by geopolitical boundaries, rather it is organic regions that share a certain quantitative value or interval scale. The regions are formed by interpolated ‘isolines’ connecting points of similar measurement to form distinct surface areas. Each area is then colour coded to represent the relevant quantitative value.

EXAMPLE
Mapping the degree of dialect similarity across the USA.

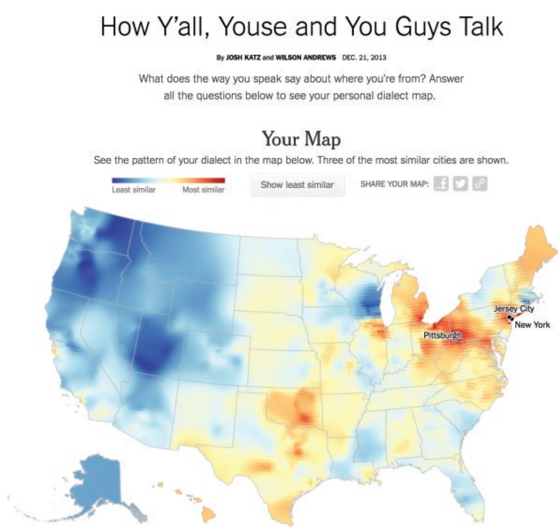


Figure 6.45 How Y’all, Youse and You Guys Talk, by Josh Katz (*New York Times*)

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be especially helpful to offer selectable tooltips to view quantitative values and category or location labels for any region on the display.

ANNOTATION: Depending on the shapes of the regions displayed, direct labelling may be limited to just a number of noteworthy values. Any colours used must be explained through the inclusion of a legend. If you choose to include a detailed map image in the background, do not include any unnecessary geographic details that add no value to the spatial orientation or interpretation (e.g. roads, building structures).

COLOUR: The outline colour and stroke width for each spatial area should be distinguishable enough to define the shape but not so prominent as to dominate. Usually, a light-grey or white-coloured stroke will suffice. Sometimes variation in pattern may be included, as well as colour, to represent values that may be uncertain or incomplete. When background map images are included, consider making them semi-transparent or light in colour to avoid competition for attention with the more important data layer.

COMPOSITION: There are many different mapping projections for spatially representing the regions of the world on a plane surface. Be aware that the transformation adjustments made by some of these projections can distort the size of regions of the world, inflating their size relative to other regions, so you will need to pick a projection that is appropriate to the spatial view you are providing.

VARIATIONS & ALTERNATIVES

There are specific applications of isarithmic maps used for showing elevation (‘contour maps’), atmospheric pressure (‘isopleth maps’) or travel-time distances (‘isochrone maps’). Sometimes you might use isarithmic maps to show a categorical status (perhaps a binary state) instead of a quantitative scale. ‘Choropleth maps’ will be the method used if your data is organised by bound regions.



PROPORTIONAL SYMBOL MAP

ALSO KNOWN AS Graduated symbol map

C H R T **S**

OVERLAYS

REPRESENTATION DESCRIPTION

A proportional symbol map displays quantitative values for locations on a map. The values are represented via proportionally sized shapes (usually circles), which are positioned with the centre mid-point over a given location coordinate. Colour is sometimes used to introduce further categorical distinction.

EXAMPLE Mapping the origin and size of funds raised across the USA for Democrat candidate Hillary Clinton during the first half of 2015.

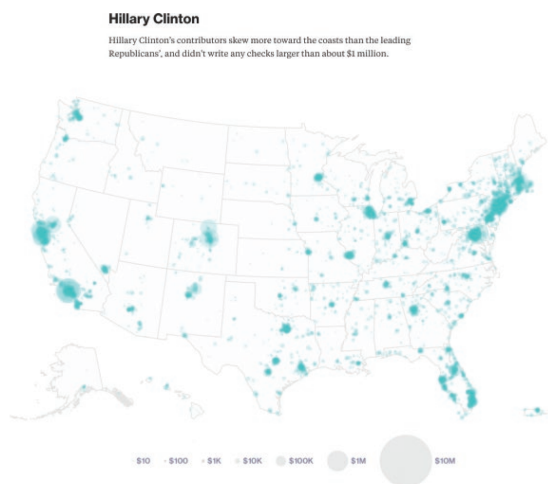


Figure 6.46 Here's Exactly Where the Candidates' Cash Came From, by Zach Mider, Christopher Cannon, and Adam Pearce (Bloomberg Visual Data)

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be especially helpful to offer selectable tooltips to view quantitative values and category or location labels for any region on the display.

ANNOTATION: Depending on the size and overlapping of shapes displayed, direct labelling may be limited to just a number of noteworthy values. Any size scales and colours used must be explained through the inclusion of a legend. If you choose to include a detailed map image in the background, do not include any unnecessary geographic details that add no value to the spatial orientation or interpretation (e.g. roads, building structures).

COLOUR: The outline colour and stroke width for each spatial area should be distinguishable enough to define the shape but not so prominent as to dominate. Usually, a light-grey or white-coloured stroke will suffice. The largest shapes may overlap, in spatial terms, with other nearby locations and sometimes even hide them completely. The use of semi-transparent colours can help avoid the effect of total occlusion. When background map images are included, consider making them semi-transparent or light in colour to avoid competition for attention with the more important data layer.

COMPOSITION: The geometric accuracy of the shape mark size calculation is paramount: it is the area you are modifying, not the diameter/radius. There are many different mapping projections for spatially representing the regions of the world on a plane surface. Be aware that the transformation adjustments made by some of these projections can distort the size of regions of the world, inflating their size relative to other regions, so you will need to pick a projection that is appropriate to the spatial view you are providing.

VARIATIONS & ALTERNATIVES

The main variations usually involve different geometric shapes being used. Alternatives include the 'choropleth map', which colour codes regions, or the 'dot map', which uses dots to represent all items across a spatial region.



PRISM MAP

ALSO KNOWN AS Isometric map, spike map, datascape

C H R T **S**

OVERLAYS

REPRESENTATION DESCRIPTION

A prism map displays quantitative values for locations on a map. The values are represented via proportionally sized lines, appearing as 3D bars, that typically cover a fixed surface area of space and are then sized through height to proportionally represent the quantitative value at each location. Attributes of colour are sometimes used to emphasise large values in particular.

EXAMPLE
Mapping the population of trees for each 180 square km of land across the globe.

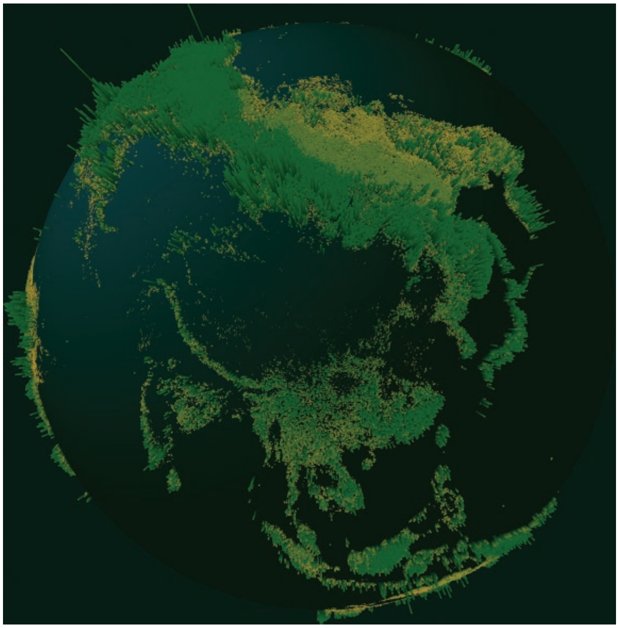


Figure 6.47 Trillions of trees, by Jan Willem Tulp

PRESENTATION TIPS

INTERACTIVITY: Ideally prism maps would be accompanied with interactive features that allow panning around the map region to offer different viewing angles that overcome the perceptual difficulties of judging the 3D presentations of data in a 2D view. Otherwise, smaller values can find themselves hidden behind larger forms, just as small buildings are hidden by skyscrapers in a city.

ANNOTATION: Direct labelling is usually impractical, so the most important feature of annotation is to indicate the size scales used in the map display. If you choose to include a detailed map image in the background, do not include any unnecessary geographic details that add no value to the spatial orientation or interpretation (e.g. roads, building structures).

COLOUR: When background map images are included, consider making them semi-transparent or light in colour to avoid competition for attention with the more important data layer.

VARIATIONS & ALTERNATIVES

Alternatives to the prism map, especially to avoid the 3D form, include the ‘proportional symbol map’, which uses proportionally sized geometric shapes, and the ‘choropleth map’, which colour codes regional shapes.

DOT MAP

ALSO KNOWN AS Dot distribution map, pointillist map, location map, dot density map

C H R T **S**

OVERLAYS

REPRESENTATION DESCRIPTION

A dot map displays the distribution of phenomena on a map. It uses point marks to plot data items at specific geographic coordinates. Items might be representative of instances of people, notable sites or incidences. The point marks are usually small circles with attributes of colour used to distinguish categorical classifications. Sometimes a dot represents a one-to-one phenomenon (i.e. a single record at that location) or one-to-many phenomena (i.e. for an aggregated statistic whereby the location represents a logical mid-point), usually depending on the potential relevance and/or sensitivity of directly plotting phenomena at precise locations.

EXAMPLE Mapping each resident of the USA based on the location at which they were counted during the 2010 Census across different ethnicities.

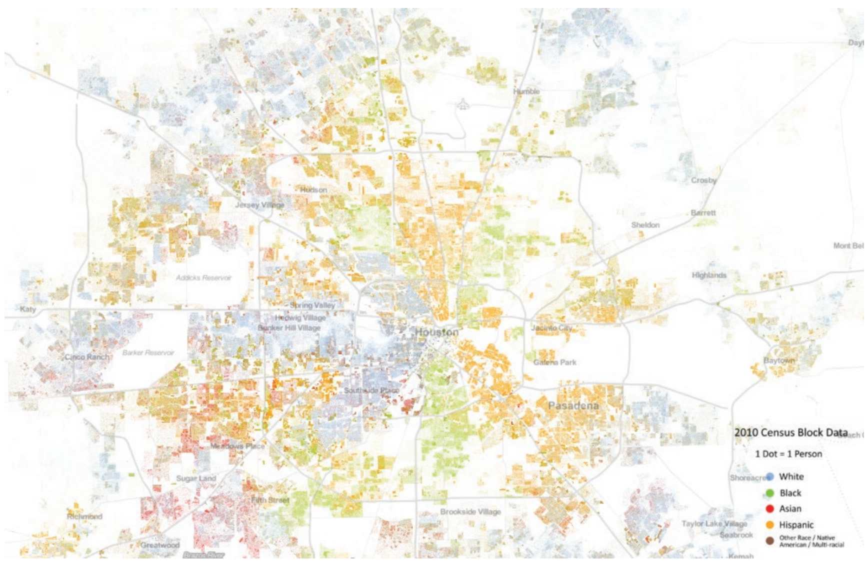


Figure 6.48 The Racial Dot Map: Image Copyright, 2013, Weldon Cooper Center for Public Service, Rector and Visitors of the University of Virginia (Dustin A. Cable, creator)

PRESENTATION TIPS

INTERACTIVITY: One method for dealing with viewing high quantities of observations is to provide interactive semantic zoom features, whereby each time a user zooms in by one level of focus, the unit quantity represented by each dot decreases, from a one-to-many towards a one-to-one relationship. Filtering options to exclude or highlight certain selections may also aid the process of understanding.

ANNOTATION: Direct labelling is rarely applied. Clear legends explaining the dot unit scale and any colour associations should ideally be placed as close to the map display as possible. If you choose to include a detailed map image in the background, do not include any unnecessary geographic details that add no value to the spatial orientation or interpretation (e.g. roads, building structures).

COLOUR: If colours are being used to distinguish the different categories, ensure these are as visibly different as possible. When background map images are included, consider making them semi-transparent or light in colour to avoid competition for attention with the more important data layer.

COMPOSITION: Dot maps should be displayed using an equal-area projection, as the precision of the plotted locations is usually paramount. From a readability perspective, try to find a balance between making the size of the dots small enough to preserve their individuality but not too tiny as to be indecipherable.

VARIATIONS & ALTERNATIVES

A 'dot density map' is a variation that involves plotting a representative quantity of dots equally (but randomly) across and within a defined spatial region. The position of individual dots is therefore not to be read as indicative of precise locations but used to form a measure of quantitative density. This offers a useful alternative to the choropleth map, especially when categorical separation of the dots through colour is of value.

FLOW MAP

ALSO KNOWN AS Connection map, route map, stream map, particle flow map

C H R T **S**

OVERLAYS

REPRESENTATION DESCRIPTION

A flow map shows the characteristics of movement or connections between phenomena across spatial regions. There is no fixed recipe for a flow map, but it generally displays characteristics of origin and destination (positions on a map), route (using organic or vector paths), direction (using arrow or tapered line width), categorical classification (colour) and quantitative measurement (line weight or, if animated, motion speed).

EXAMPLE Mapping the average number of vehicles using Hong Kong's main network of roads during 2011.

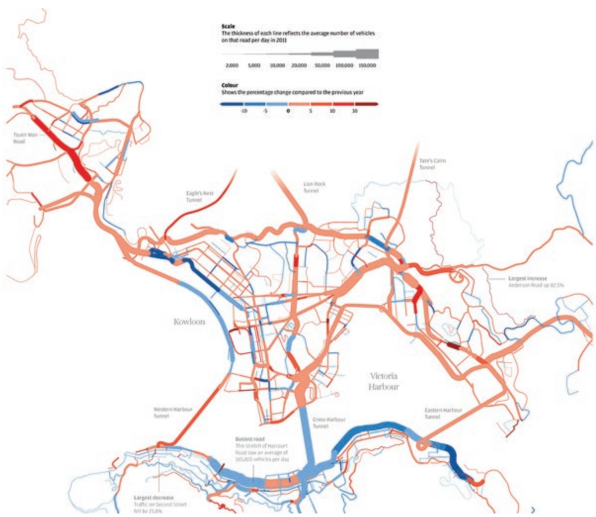


Figure 6.49 Arteries of the City, by Simon Scarr (*South China Morning Post*)

PRESENTATION TIPS

INTERACTIVITY: Animated sequences may provide a useful presentation method when the phenomena are characteristic of some notion of movement.

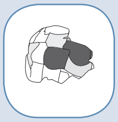
ANNOTATION: Annotation needs will be unique to each approach and the inherent complexity or otherwise of the display. Often the general patterns may offer the sufficient level of readability without the need for imposing amounts of value labels, but clear legends explaining the associations with any attributes used will be important to include. If you choose to include a detailed map image in the background, only include any relevant geographic details that offer spatial orientation or interpretation to the nature of flow being represented (e.g. roads, rivers, oceans).

COLOUR: If colours are being used to distinguish the different categories, ensure these are as visibly different as possible. When background map images are included, consider making them semi-transparent or light in colour to avoid competition for attention with the more important data layer.

COMPOSITION: Some degree of geographic distortion or smoothing of flow routes may be required. Decisions about the degree of interpolation applied to line smoothing or the merging of relatively similar pathways may be entirely legitimate, but ensure that this is made clear to the viewer. There are many different mapping projections for spatially representing the regions of the world on a plane surface. Be aware that the transformation adjustments made by some of these projections can distort the size of regions of the world, inflating their size relative to other regions, so you will need to pick a projection that is appropriate to the spatial view you are providing.

VARIATIONS & ALTERNATIVES

There are several variations for how you might label different applications of displaying flow. It generally depends on whether you are showing point A to point B journeys ('connection maps'), more intricate pathways ('route maps') or organic phenomena ('particle flow maps').



AREA CARTOGRAM

ALSO KNOWN AS Contiguous cartogram, density-equalising map

C H R T **S**

DISTORTIONS

REPRESENTATION DESCRIPTION

An area cartogram displays the quantitative values associated with distinct, definable spatial regions on a map. Each geographic region is represented by a polygonal area based on its outline shape with the collective regional shapes forming the entire landscape. Quantitative values are represented by proportionately distorting (inflating or deflating) the relative size of and, to some degree, shape of the respective regional areas. Traditionally, area cartograms strictly aim to preserve the neighbourhood relationships between different regions. Attributes of colour are often used to represent the quantitative measurements and/or to associate the region with a categorical classification. Area cartograms require the reader to be relatively familiar with the original size and shape of regions in order to be able to establish the degree of relative change in their proportions.

EXAMPLE Mapping the measures of climate change responsibility compared with vulnerability across all countries.

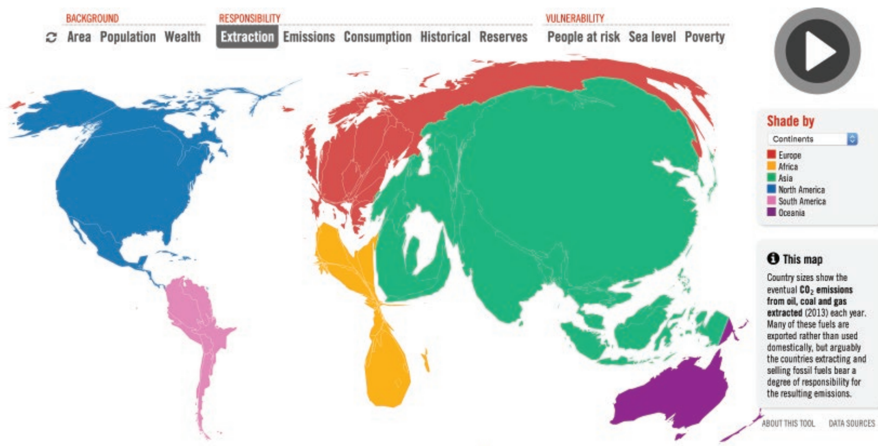


Figure 6.50 The Carbon Map, by Duncan Clark and Robin Houston (Kiln)

PRESENTATION TIPS

INTERACTIVITY: Animated sequences enabled through interactive controls can help to better identify instances and degrees of change, but usually only over a small set of regions and only if the change is relatively smooth and sustained. Manual animation will help to provide more control over the experience. Selectable tooltips to view quantitative values and category or location labels for any region on the display may also prove useful.

ANNOTATION: Directly labelling the regional areas with geographic details and the value they hold is likely to lead to too much clutter. As it is difficult to assess the degree of distortion and, indeed, often to identify the regions themselves, it can be useful to present a thumbnail view of the undistorted original geographic layout to help readers orient themselves with the changes. Additionally, a limited number of regional labels might be included to provide direct spatial context and orientation. Any colours used must be explained through the inclusion of a legend.

COLOUR: The outline colour and stroke width for each spatial area should be distinguishable enough to define the shape but not so prominent as to dominate. Usually, a light-grey or white-coloured stroke will suffice.

VARIATIONS & ALTERNATIVES

Unlike contiguous cartograms, non-contiguous cartograms tend to preserve the shape of the individual polygons but modify the size and the neighbouring connectivity to other adjacent regional polygon areas. The best alternative ways of showing similar data would be to consider using the 'choropleth map' or 'Dorling cartogram'.

DORLING CARTOGRAM

ALSO KNOWN AS Demers cartogram

REPRESENTATION DESCRIPTION

A Dorling cartogram displays the quantitative values associated with distinct, definable spatial regions on a map. Each geographic region is represented by a circular mark which is proportionally sized to represent a quantitative value. The placement of each circle loosely resembles the region's geographic location with general preservation of neighbourhood relationships between adjacent shapes. Attributes of colour hue are often used to associate each spatial region with a categorical classification.

EXAMPLE
Mapping the share of people using the Internet by country as at 2015.

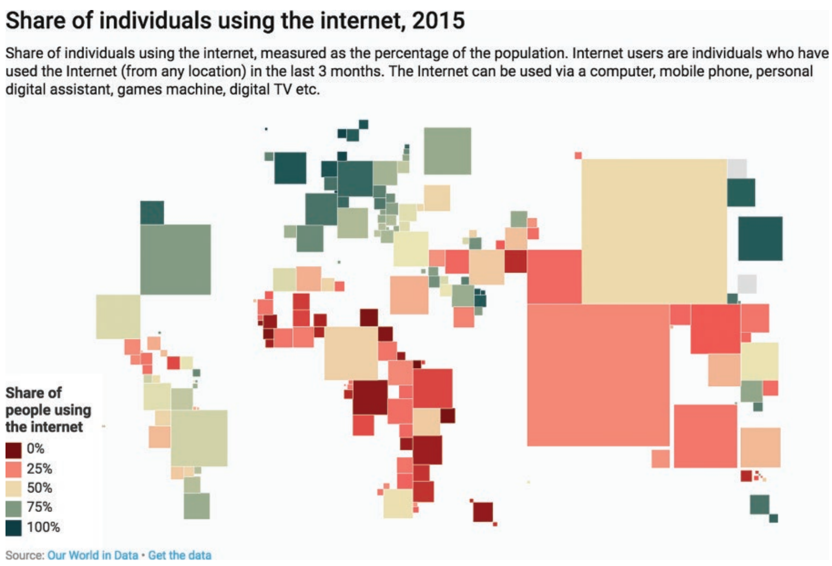


Figure 6.51 Share of Individuals Using the Internet, 2015, by Lisa Rost

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be helpful to offer selectable tooltips to view quantitative values and category or location labels for any region on the display.

ANNOTATION: Directly labelling the shapes with geographic details and the values they hold is common, though you might restrict this to only circles that are of sufficient size to hold such annotations. Any colours used must be explained through the inclusion of a legend.

COMPOSITION: Preserving the layout adjacency with neighbouring regions is important. Dorling cartograms tend not to allow circles to overlap or occlude, so some accommodation of large values might result in some location distortion.

VARIATIONS & ALTERNATIVES

A variation on the approach, called the 'Demers cartogram', involves the use of rectangular marks instead of circles. This offers an alternative way of connecting adjacent shapes. Other alternative chart types to consider would be the 'area cartogram' or the 'choropleth map'.

GRID MAP

ALSO KNOWN AS Cartogram, bin map, equal-area cartogram, hexagon bin map

C H R T **S**

DISTORTIONS

REPRESENTATION DESCRIPTION

A grid map displays the quantitative values associated with distinct, definable spatial regions on a map. Each geographic region (or a statistically consistent interval of space, known as a ‘bin’) is represented by a fixed-size uniform shape, sometimes termed a **tile**. The marks used tend to be squares or hexagons, though any tessellating shape might help to arrange all regional tiles into a collective shape that roughly fits the real-world geographic adjacency. Attributes of colour are applied to each regional tile either to represent a quantitative measurement or to associate the region with a categorical classification.

EXAMPLE Showing the percentage of people voting to leave and remain across the UK electoral seats during the EU referendum in 2016.

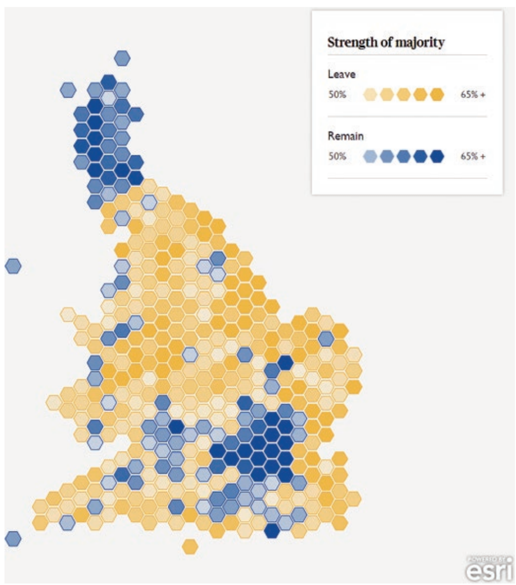


Figure 6.52 Share of People Voting to Leave and Remain During the EU Referendum in 2016, by Ben Flanagan

VARIATIONS & ALTERNATIVES

‘Hexagon bin maps’ are specific deployments of the grid map that offer a layout formed by a high resolution of smaller hexagons to preserve localised details.

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be helpful to offer selectable tooltips to view quantitative values and category or location labels for any region on the display.

ANNOTATION: Directly labelling the shapes with geographic details is usually impractical due to the small size of each point mark, unless short abbreviated values can suitably represent the location label. Legends explaining the colour associations must be included.

COMPOSITION: The main composition challenge is to determine the right geographic level for each constituent tile to be representative of, and to optimise, the best-fit collective layout that preserves as many neighbouring relationships as possible.

6.2 Influencing Factors and Considerations

You have now been through the gallery of chart-type options learning more about their specific roles and what design features may enhance their particular deployment. Even if you have a fairly clear idea about which chart(s) you might choose, there are other factors that may influence your final decision of how to represent your data. There is a blend of considerations to draw from your progress through the first three preparatory stages of the design process, supplemented by the enduring need to satisfy the three principles of good visualisation design, as presented in Chapter 2.

Technological: What charts you can actually make and how easily you can personally create them is a big factor. Data visualisation technologies offer different chart-making capabilities and it can be hard navigating through the options that exist. To assist with this, you might consider consulting the 'Chartmaker Directory' (chartmaker.visualisingdata.com/). This digital resource organises a huge catalogue of useful references that will offer an answer to the most common of questions: 'Which tool do you need to make that chart?'

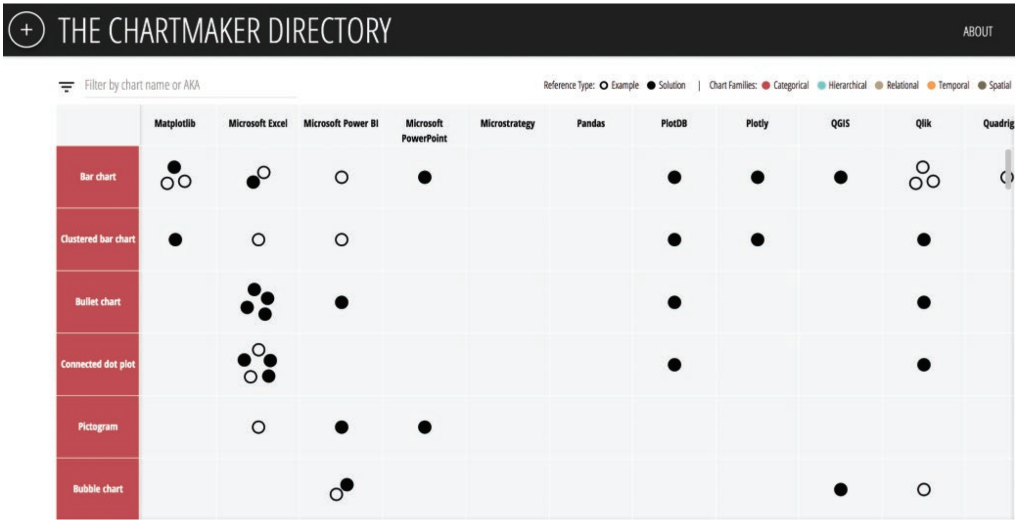


Figure 6.53 Screenshot of the 'Chartmaker Directory'

The directory's content is presented through a tabular layout (Figure 6.53). Across the top of the table are a selection of around 40 chart-making tools. A comprehensive list of different chart types is presented down the side matching the gallery you have just explored. Inside the intersecting cells, you will find unfilled and filled circular markers representing a reference in the directory:

- An *unfilled* mark represents a link to an example, providing evidence that a given chart can be made in a given tool. Read it as 'here's a link to a bar chart made using Excel', for example.

- A *filled* mark represents a link to a solution, which provides guidance on how to create a given chart with a given tool. Solutions might exist as 'how-to' tutorials with step-by-step instructions, video demonstrations, downloadable workbooks/templates or reusable code.

The directory is constantly growing as more chart-making solutions and examples are discovered for each tool. In particular, many valuable references present smart workarounds or 'out of the box' thinking that employ unconventional techniques to create a chart in a tool that normally would not seem possible.

> 'The capability to cope with the technological dimension is a key attribute of successful students: coding – more as a logic and a mindset than a technical task – is becoming a very important asset for designers who want to work in Data Visualisation. It doesn't necessarily mean that you need to be able to code to find a job, but it helps a lot in the design process. The profile in the (near) future will be a hybrid one, mixing competences, skills and approaches currently separated into disciplinary silos.'
> **Paolo Ciuccarelli, discussing students on his Communication Design Master Programme at Politecnico di Milano**

Purpose: Having the technical ability to create a broad repertoire of chart types is the *vocabulary* of this discipline; judging when to use them is the *literacy*. The first question to ask yourself is if you even need to represent your data in chart form. Will this enable new qualities and relationships in your data to be seen? Do not rule out the value of a table if providing a means for your viewer to look up and reference values. It might be a more suitable solution option.

This brings us back to the discussion about the importance of defining the tone of your project. Are you aiming to facilitate the reading of data or the feeling of data? Is it more important to offer precise value judgements or should more emphasis be placed on general sense-making about the big, medium and small values? Were there emotional qualities you wanted to emphasise or suppress?

In his book *Semiology Graphique*, published in 1967, Jacques Bertin proposed the idea that different ways of encoding data might offer varying degrees of accuracy in the perception of data values. In 1984, William Cleveland and Robert McGill published a seminal paper, 'Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods'. This offered more empirical evidence of Bertin's thoughts. From this study they developed a general ranking that explained which attributes used to encode quantitative values would facilitate the highest degree of perceptual accuracy. In 1986, Jock Mackinlay's paper, 'Automating the Design of Graphical Presentations of Relational Information', further extended this to include proposed rankings for encoding categorical (nominal and ordinal) data, as well as quantitative values. The table shown in Figure 6.54 presents the 'Ranking of Perceptual Tasks'.

What this ancestry of studies reveals is that the use of certain attributes to encode certain types of data may make it quicker, easier and more accurate to judge the values portrayed. Two classic illustrations of this notion are shown below. Looking at Figure 6.55, if A is 10, how big is B?

Qualitative Nominal	Qualitative Ordinal	Quantitative Interval, Ratio
Position	Position	Position
Colour (Hue)	Pattern (Density)	Size (Length)
Pattern (Texture)	Colour (Lightness)	Angle
Connection	Colour (Hue)	Size (Area)
Containment	Pattern (Texture)	Size (Volume)
Pattern (Density)	Connection	Pattern (Density)
Colour (Lightness)	Containment	Colour (Lightness)
Symbol	Size (Length)	Colour (Hue)
Size (Length)	Angle	Pattern (Texture)
Angle	Size (Area)	Connection
Size (Area)	Size (Volume)	Containment
Size (Volume)	Symbol	Symbol

Note that the attribute of 'Motion' was not included in this study. For the purposes of this display, 'Angle' and 'Slope' are combined whereas they were distinguished as separate in the study.

Figure 6.54 The Ranking of Perceptual Tasks, adapted from Mackinlay (1986)

In both cases the answer is B equals 5. Although B in the bar chart being of size 5 feels about right, the idea that circle B is also 5 feels less so. Our visual system is superior in its accuracy when performing relative judgements for a line, in comparison with a shape. This is explained by the fact that judging the variation in size of lines involves detecting change in a linear dimension (length), whereas the variation in size of a geometric shape like a circle happens across a quadratic dimension (area). If you look at the rankings in Figure 6.54 in the 'Quantitative' column, you will see the encoding attribute of *Length* is ranked higher than the attribute of *Area*.

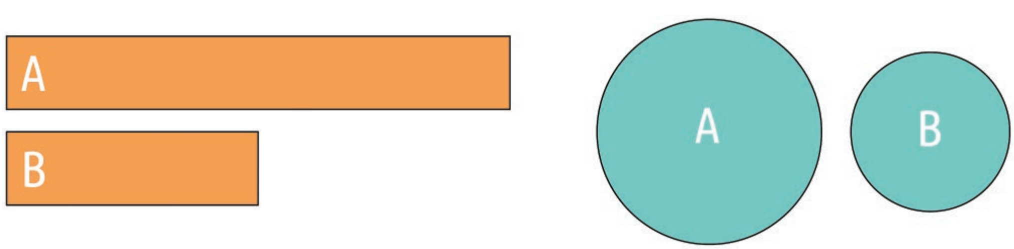


Figure 6.55 Comparison of Judging Line Size vs Area Size

Now let's consider a demonstration of perceptual accuracy when using different dimensions of colour variation to represent nominal values. In the charts shown in Figure 6.56 you can see that different attributes are used to represent the categorical groupings in the two scatter plots.

On the left you see variation in the attribute of colour hue (blue, orange and green) to classify the distinct categories; on the right you see the attribute of symbol (diamond, circle and square) applied similarly.

What you will find is a more immediate, effortless and accurate experience in identifying the groupings of the coloured category markers compared with the symbol-based equivalents. It is easier to observe classifications through variation in colour than it is using variation in symbol, as supported by *colour hue* being ranked higher than *shape* for nominal data types as shown in the table in Figure 6.54.

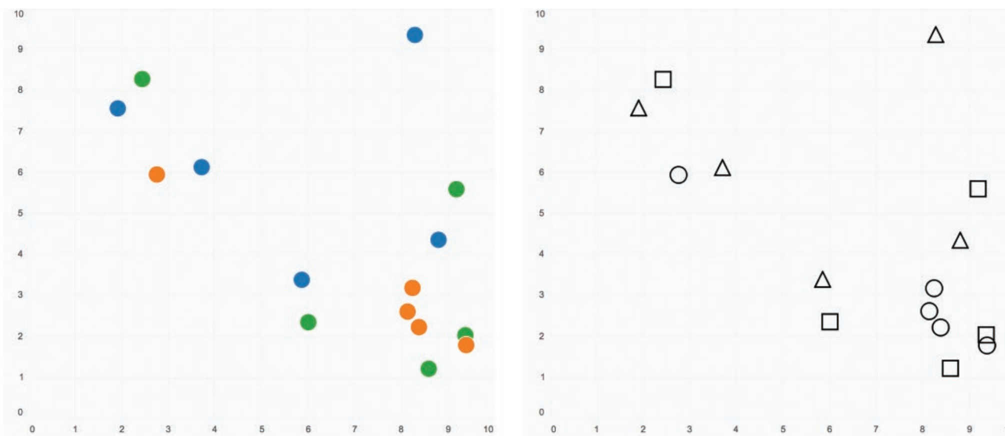


Figure 6.56 Comparison of Judging Categorical Associations Using Variation in Hue vs Variation in Shape

You can see from these simple demonstrations that there are clearly ways of encoding data that will make it easier to read values accurately and efficiently. However, as Cleveland and McGill stress in their paper, this should only be taken as guidance, commenting that the ranking of attributes 'does not result in a precise prescription for displaying data but rather is a framework within which to work'.

This is important to acknowledge because you have to weigh up whether precise perceiving is actually what you wish to offer your viewers. As stated in Chapter 3, sometimes getting the 'gist' of data values is sufficient. You might therefore determine that selecting a chart that uses the attribute of size through variation in area, which is lower down the quantitative attribute rankings, offers a suitable balance. Judging the hierarchy of large, medium and small features may be sufficient for your needs. It depends on your purpose.

Sometimes, you will have scope in your encoding choices to incorporate a certain amount of visual immediacy in accordance with your topic. I warned earlier about the need to be driven by your data and not by your ideas, but sometimes there is scope to squeeze out extra stylistic associations between the visual and the content. The flowers of the Better Life Index feel consistent in metaphor with the idea of better life: the more in bloom the flowers, the more colourful and prouder each petal appears and the better the quality of life in that country.



Figure 6.57 Simulated Dendrochronology of U.S. Immigration, by Pedro Cruz, John Wihbey, Avni Ghael and Felipe Shibuya

'I've come to believe that pure beautiful visual works are somehow relevant in everyday life, because they can become a trigger to get people curious to explore the contents these visuals convey. I like the idea of making people say "oh that's beautiful! I want to know what this is about!" I think that probably (or, at least, lots of people pointed that out to us) being Italians plays its role on this idea of "making things not only functional but beautiful".' **Giorgia Lupi, Co-founder and Design Director at Accurat**

What you are trying to represent may not be possible using a conventional chart. Another example that draws from nature is shown in Figure 6.57. This piece uses the notion of dendrochronology – the study of tree rings – to create a compelling portrayal of the history of immigration into the USA. Each ring is a decade container, working outwards through chronological decades. Within each container ring are dots corresponding to 100 immigrants. The colours indicate the origin continents or major regions. The outcome is a stunning concept that perfectly aligns subject matter and visual encoding.

Data type and shape: The types of data and range of values you are trying to display will have a bearing on which charts you can use, and, of those, which will best portray what you want to say. Any chart type will only accommodate certain types of data. For example, if you want to use a line chart, you will need one or more continuous series of quantitative values that have a dimension of temporal data. Additionally, the viability of any chart choice will be determined by how well it accommodates the range of values you wish to include. As ever, this depends on what it is you want to say.

Let's suppose you are producing some simple analysis about the market share of browsers. The first chart you consider is the pie chart. To use this you will need quantitative values, in the form of percentages, for different categories that aggregate to a true 'whole' (nothing more, nothing less, than 100%). The data shows there is a market share breakdown across 30 discrete browsers.

As you can see in Figure 6.58, there are a few issues with the pie chart (A). If it is important for a viewer to judge values for each of the 30 browsers with a certain degree of accuracy, the pie chart will not be fit for purpose. It gets harder to perceive the size of each slice after the first three or four. Furthermore, the colour associations as shown in the legend are indiscernible. We need a plan B. In this case you might switch to a bar chart. Even though this chart belongs to a different 'family' you can still use it to represent parts-of-a-whole percentages for each browser item. This will offer an improved option to make it more readable, both in judging the values and through the proximity of the category labels to each bar. It does, though, result in a lot of empty space due to the skewed shape of the data values.

If you are really seeking to enable the readability of each value, you may try to convey just how dominant Chrome is as one part of this whole. You might therefore revert to using a pie chart (C) to include all the discrete browser categories, but label only the Chrome part and summarise the rest as a single 'All others' value. You only need the Chrome value to be seen as 'biggest' compared with the many other competitors battling for but losing out on the dominant market share. If the visibility of the many other parts is not important, group them into a single 'All others' value so now you have a simple two-slice pie (D) or a donut chart (E) if you wish to exploit the empty centre to accommodate the summary value labels.

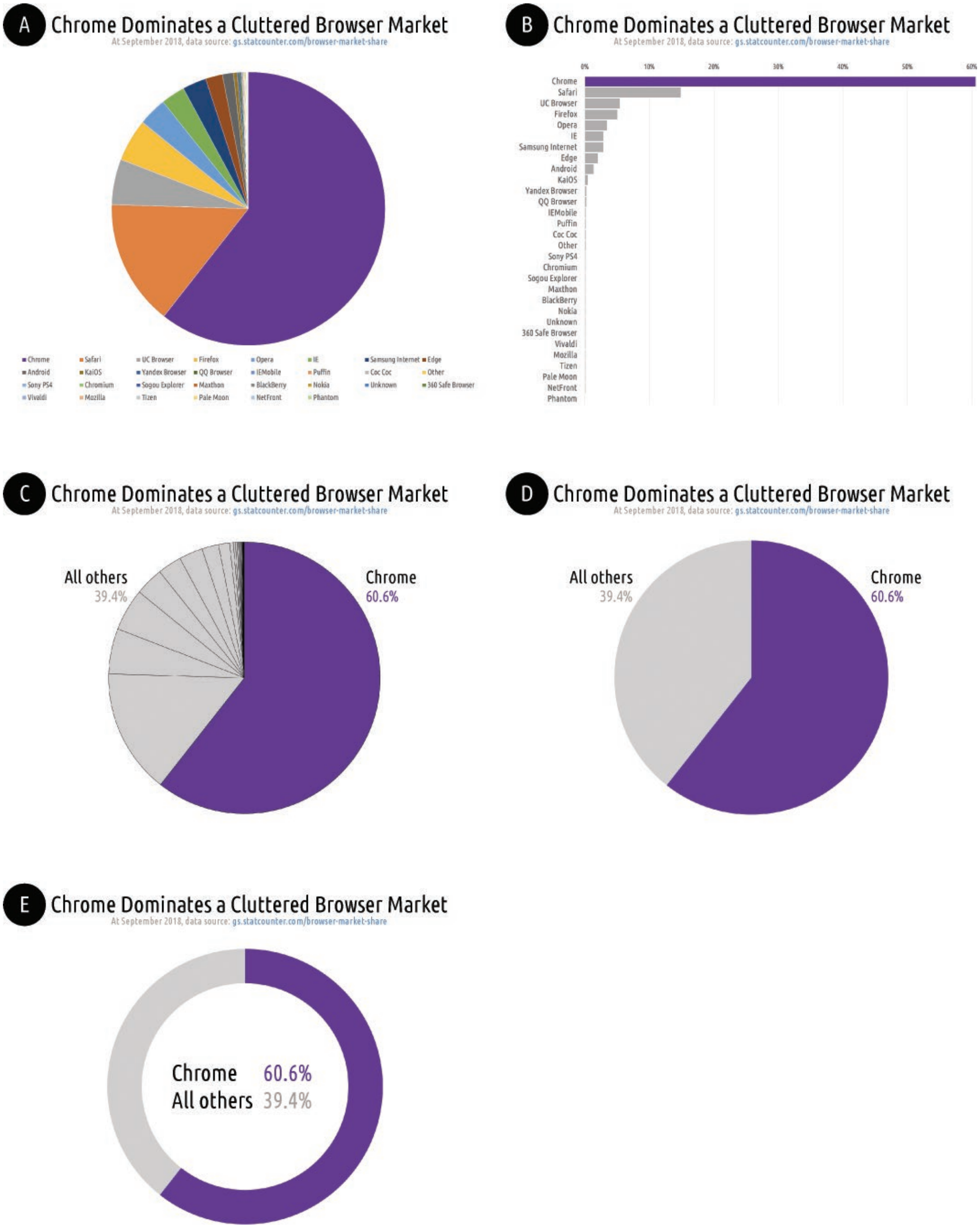


Figure 6.58 Iterations of Different Chart Options to Show the Same Data

This illustration demonstrates how you only know if a chart will serve your purpose once you try it out with real data. After that, consider variations in chart and/or transformations of your data to find the best way to show what you really want to say.

Data exploration: One consistently useful pointer to how you might visually communicate your data is to consider which techniques helped *you* to unearth key insights when you were

visually exploring the data. What chart types have you already tried out and maybe found to reveal interesting patterns? Exploratory data analysis, in many ways, offers this bridge to visual communication: the charts you use to see data for yourself often represent prototype thinking about how you might communicate real data to others. The way you style the chart may differ, but if a method is already working, why not utilise the same approach again?

Editorial angle: When defining your editorial angle(s) you are expressing what specific aspect of understanding you are attempting to portray to your viewers. This helps you to determine which chart type might be most relevant or at least which family across the CHRTS taxonomy will provide the best option to pick from. Always give yourself time to spend on the editorial stage, carefully articulating *what* you want to say before you get too carried away with picking *how*.

Trustworthy design: In the discussion about tone I explained how you might sacrifice precision in the perception of values to suit the purpose of your work. Precision in perception is one thing, but precision in design is a different matter and one for which there should be no compromise. Being accurate in your portrayal of data is a fundamental obligation. There are many ways in which viewers can be deceived through incorrect and inappropriate encoding choices, whether they are intended or not.

Geometric miscalculations are a common mistake. When using the area of shapes to represent different quantitative values, the underlying geometry needs to be calculated accurately. For example, using circular shapes to show a quantitative value of 20 compared with another of 10, you would just half the diameter of the second, right? Wrong.

The illustration in Figure 6.59 shows the incorrect and correct ways of encoding two quantitative values through circle size, where value A is twice the size of B. The orange circle for B has half the *diameter* of A, the green circle for B has half the *area* of A. Using variation in diameter distorts the perceived size of circle B as being far smaller than the value actually is. Viewers base estimates of quantitative size through the area of a circle, not its diameter. Therefore, the green circles demonstrate the correct way to encode these values.

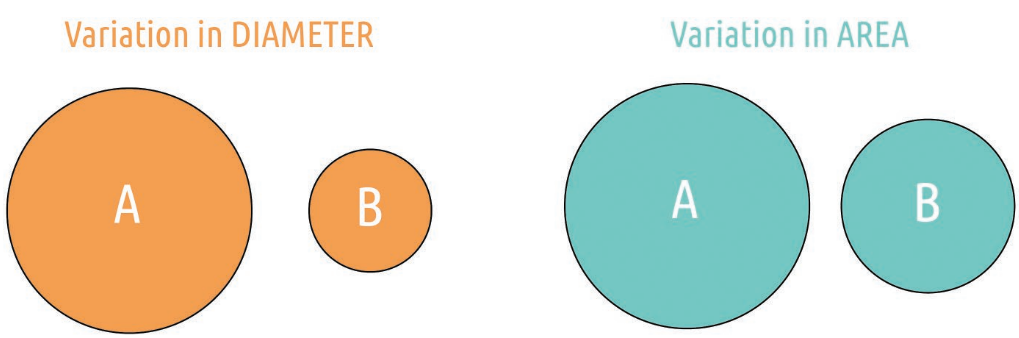


Figure 6.59 The Correct and Incorrect Way to Encode Variation in Shape Size

Another representation accuracy issue causing problems for size judgements concerns truncated axis scales. When quantitative values are encoded through the height or length of size (e.g. for bar charts), truncating the value axis (not starting the range of quantitative values from the

origin of zero) distorts the size judgements. I will revisit this issue in Chapter 10 because it is ultimately a consideration about the sizing of chart-scale ranges, which I deem to be a matter of composition.

Another design issue that can distort data is 3D decoration. In the majority of cases, the use of 3D charts is, at best, unnecessary and, at worst, hugely distorting. Though I concede that there can be a certain appeal to the physical appearance of 3D charts, it is not an effective choice for trustworthy practices. It is often seen applied to a chart when the visualiser is motivated by a desire to demonstrate technical competence with a tool or encouraged by stakeholders who want to see charts made to look 'fancy' or 'cool'.

Using psuedo-3D decoration, when you have only two dimensions of data, is gratuitous and will distort the viewer's ability to judge values with any degree of acceptable accuracy. As illustrated in Figure 6.60, when forming value estimates of the angles and sectors in the respective pie charts, the 3D version makes it much harder to form accurate judgements. The tilting of the isometric plane amplifies the front part of the chart and diminishes the back. It also introduces a raised 'step' which is purely decorative, thus embellishing the judgement of the sector sizes.

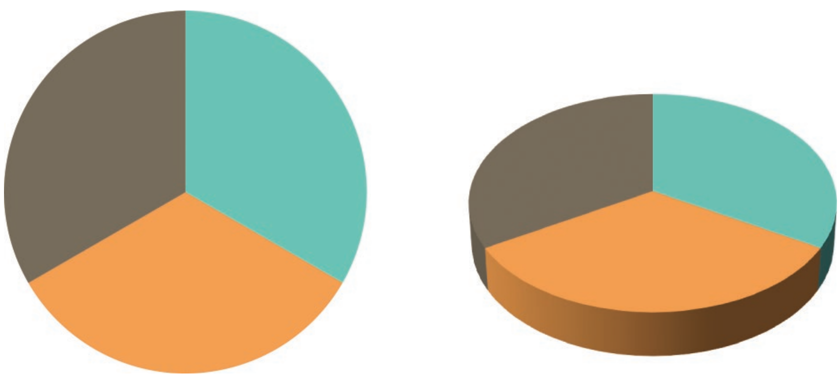


Figure 6.60 Illustrating the Distortions Created by 3D Decoration

For charts genuinely based on three dimensions of data, a 3D representation should only be considered reasonable if the viewer is provided with the means to adjust the field of view. This will help to overcome the distortion of distance and perspective, creating multiple potential 2D viewing angles. 'All the Buildings in Manhattan', Figure 6.61, offers a slick interactive experience that lets users navigate around a 3D view of New York City to observe the size of the buildings around Manhattan. This means you can change your field of view to determine properly the height of the modelled building shapes and make comparisons across the city.

Another legitimate application of 3D visualisation is through the potential of physical displays, perhaps using 3D printing techniques, as demonstrated by the piece shown in Figure 6.62. This portrays trajectories for every home run scored by Kris Bryant of the Chicago Cubs during 2017, including the height, distance and landing position of each shot.

The final matter related to trustworthiness concerns thematic mapping, specifically the often contentious matter of choosing a map projection. The Earth is not flat. Although advances

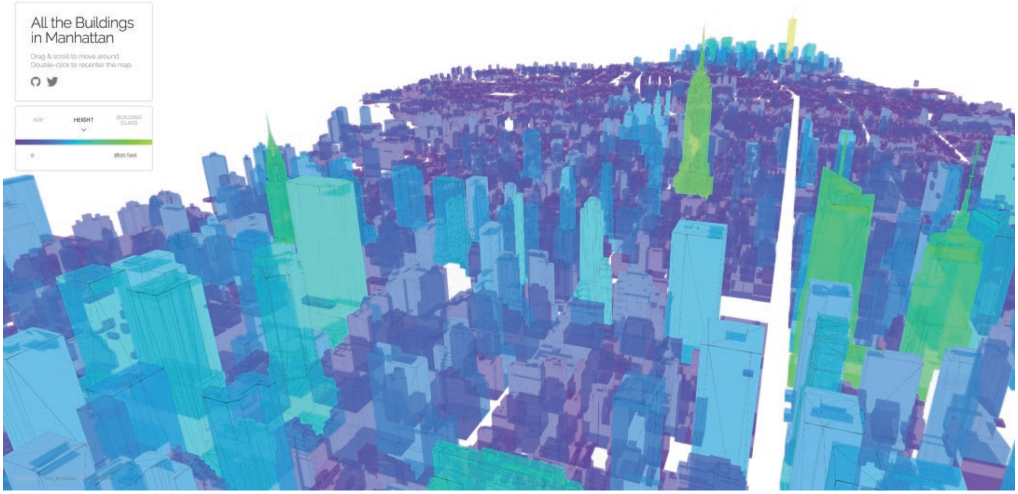
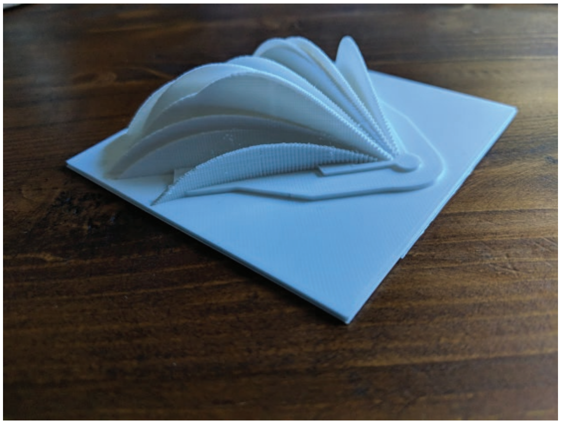


Figure 6.61 All the Buildings in Manhattan, by Taylor Baldwin (tbaldwin.net, @taylorbaldwin)

Figure 6.62 Representing Three Dimensions of Data (Baseball Home Run Trajectories) in a 3D Space



in technology are enabling interaction with 3D portrayals of the Earth within a 2D space, the dominant form through which maps are presented portrays the Earth as a flat surface. Features such as size, shape and distance can be measured accurately on Earth, but when projected onto a flat surface a compromise has to occur. Only some of these qualities can be preserved and represented accurately. Although there are exceptionally complicated calculations attached to each spatial projection, the main features most of us need to know about are that:

- every type of map projection has some sort of distortion;
- the larger the area of the Earth portrayed as a flat map, the greater the distortion;
- there is no single right answer – it is often about choosing the least-worst case.

Thematic mapping (as opposed to mapping spatially for navigation or reference purposes) is generally best carried out using mapping projections based on 'equal-area' calculations (so the sacrifice is more on the shape, not the size). This ensures that the phenomena per unit – the values you are typically plotting – are correctly represented by proportion of regional area. For choosing the best specific projection, in the absence of perfect, the decision is usually based on which one will distort the spatial truth the least given the level of mapping required. There are many variables in play, however, based on the scope of view (world, continent or country/ sub-region), the potential distance from the equator of your region of focus and whether you

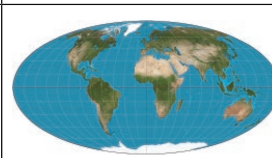
Mercator While the Mercator has been widely discredited in its role as a means of portraying the world (due to the vast distortions at the poles) it is still the most common projection found in mapping tools (where it is often termed Web Mercator). This is largely because of its rectangular dimensions that support seamless zooming. If you are determined to use this projection, you should not use it for a global view; stick to a lower regional level so the distortions are minimised, especially for regions around the equator.	
Equal Earth The Equal Earth map projection is an equal-area pseudo-cylindrical projection for world maps. It was developed in order to create a world map showing continents and countries at their true sizes relative to each other.	
Lambert Azimuthal Equal-area This spherical projection is most commonly recommended for hemisphere- or continent-level views. The European Environment Agency, for example, recommends its usage for any European mapping purpose.	
Winkel–Tripel Most of the important people who are far better informed about mapping projections than I often describe the Winkel–Tripel projection as being one of the best choices for viewing the world. Indeed, it represents the modern standard world map adopted by National Geographic.	
Mollweide In contrast to the Winkel–Tripel, the Mollweide (equal-area) projection offers greater emphasis on the accuracy of ocean areas and can be useful for atmospheric mapping (e.g. flight paths).	

Figure 6.63 A Selection of Commonly Deployed Mapping Projections. Images from Wikimedia Commons published under the Creative Commons Attribution-Share Alike 3.0 Unported Licence

are focusing on land, sea or sky (atmosphere), to name but a few. As with many other topics in this field, a discussion of mapping projections requires a dedicated text, but let me at least offer a brief outline of five different projections (Figure 6.63).

Summary: Data Representation

Visual Encoding and Charts

This chapter introduced the act of visual encoding, the fundamentals of how you represent data visually. All charts are based on a combination of marks and attributes:

- Marks: Visual placeholders representing data *items*, such as distinct records or discrete groupings.
- Attributes: Variations in the visual appearance of marks to represent the values associated with each data item.

Expanding on this introduction, you were then introduced to a wide gallery of chart types, including profiles of 49 distinct approaches, to give you a sense of the common options that exist. The charts were organised into five family groupings, based on what each type is primarily used to show:

- Categorical: Comparing categories and distributions of quantitative values.
- Hierarchical: Revealing part-to-whole relationships and hierarchies.
- Relational: Exploring correlations and connections.
- Temporal: Plotting trends and intervals over time.
- Spatial: Mapping spatial patterns through overlays and distortions.

Influencing Factors and Considerations

If these were the options, how did you make your choices? The influencing factors included:

- Technological: What charts can you make and how efficiently?
- Purpose: What is the intended 'tone' of voice your representation should convey? Where is the emphasis between reading and feeling data?
- Data type and shape: The types of data and range of values you are trying to display will have a bearing on which charts you can use.
- Data exploration: What charting methods did you use to explore your data and did any of those represent possible means for communicating to your audience?
- Editorial angle: What is the specific angle of enquiry that you want to portray visually? Is it relevant and representative of the most interesting analysis of your data?
- Trustworthy design: Avoid deception through mistaken geometric calculations, 3D decoration, truncated axis scales, corrupt charts.

General Tips and Tactics

- Do not arrive at this stage with fixed, preconceived ideas about wanting to use certain chart types: be driven by your data and by your editorial thinking.
- Do not be precious: acknowledge when you have made a wrong call or gone down a dead end.

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

7

Interactivity

In the previous chapter we explored a wide range of different options for representing data visually and learnt about the factors informing your choices. In this chapter we move on to the second element of design thinking concerning the potential features of interactivity.

It is not much more than a generation ago that most visualisations would have been created exclusively for print. The advancement of technology has now entirely altered the nature of how visualisations are produced, shared and consumed. The capabilities of modern devices and proliferation of high-speed web access have created a particularly fertile landscape for talented developers to produce engaging interactive experiences.

Not everything can, will or should be interactive. The careful judgements that characterise this entire visualisation design process are especially important when handling this layer of anatomy. Features of interactivity must be fundamentally justifiable. They must enhance and not obstruct the facilitation of understanding.

In this chapter we will temporarily switch labels from 'viewer' to 'user' as this implies a more active role for discussing interactivity. For a user to become active, there needs to be sufficient reward for making the effort. Moreover, the visualiser must ensure that offering interactivity does not reflect an abdication of responsibility. Do not pass on to the user the task of discovering insights if a context necessitates the provision of an explanatory experience. This often betrays a certain lack of commitment to editorial clarity.

However, when the circumstances are appropriate, incorporating features of interactivity into your visualisation can offer several advantages:

- It expands the physical limits of what can be consumed in a given space.
- It broadens the variety of analysis to serve different curiosities within a single project.
- It facilitates manipulations of data to accommodate varied interrogations.
- It amplifies the overall control and potential customisation of an experience.
- It increases the range of techniques for engaging users with dynamic displays.

Before deciding what features of interactivity you *will* use, we will first consider what you *could* use. We will explore some of the common methods employed to equip you with the means for interrogating, manipulating and navigating through rich digital experiences. As we

encounter each option, it will be useful for you to understand the distinctions between an *event*, the *control* and the *function*:

The *event* is the user action, such as a single click of a mouse.

The *control* is the feature to which the event action is applied, such as a dropdown menu.

The *function* is the operation that is performed, such as selecting a category.

7.1 Features of Interactivity

Filtering

The first set of techniques (Table 7.1) enables users to specify what data they wish to include or exclude from a chart display. This action effectively modifies the editorial 'framing' perspective for the current view of data.

Table 7.1 Features of Interactivity to Facilitate Filtering

Example events and controls	Example functions
Select a button or link Select an item from a menu list Select multiple items from a check-box or menu list Alter the state of a toggle or radio button Alter the position of a handle along a scale slider Alter the position of two handles along a scale slider (to create a range) Enter a value into an input box	Apply a categorical data filter (one or several combinations) Apply a quantitative data filter (one value or range of values) Reset all values to their original state

The first example in Figure 7.1 shows an excerpt from 'The Pursuit of Faster', a visualisation project looking at the evolution of result times throughout the history of the Summer Olympics. By using the categorical menus at the top of the screen, users can modify the selection of which sport and event results are currently displayed in the chart. Further categorical filters can then be applied to show or hide results for different genders and for individual medal series.

'How Nations Fare in PhDs by Sex', shown in Figure 7.2, explores the under-representation of women among the workforce of science and engineering organisations. This gender disparity is revealed by looking at the PhDs awarded to men and to women across different countries and subject areas. The interactive options provided in the dropdown menu let users select and apply different subject filters to help reveal different patterns of disparity.

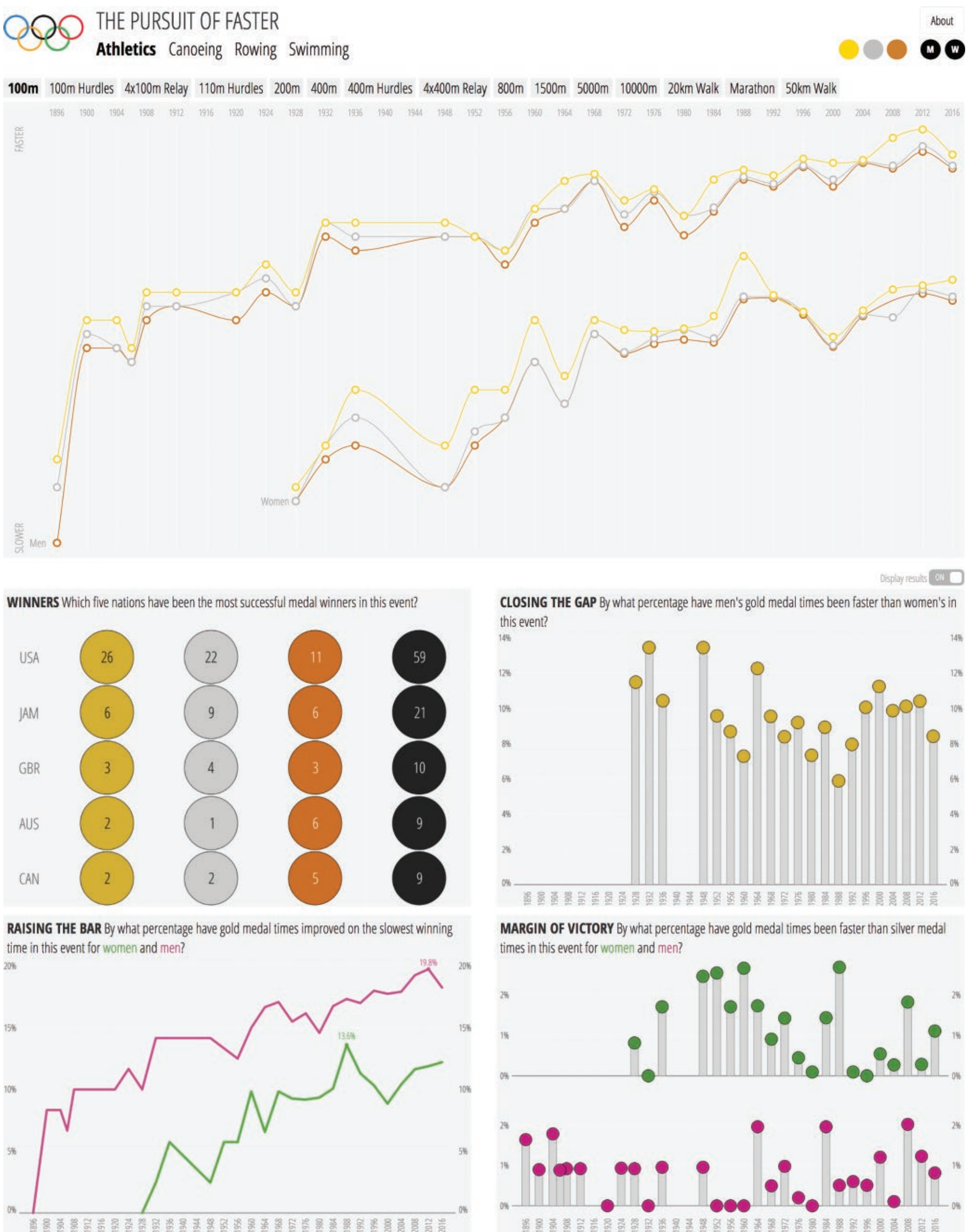


Figure 7.1 The Pursuit of Faster, by Andy Kirk and Andrew Witherley

Shown in Figure 7.3 is a census of the prevalence of species of trees found around the boroughs of New York City. This initial big-picture view creates a beautiful tapestry made up of a wide range of different tree populations across the region.

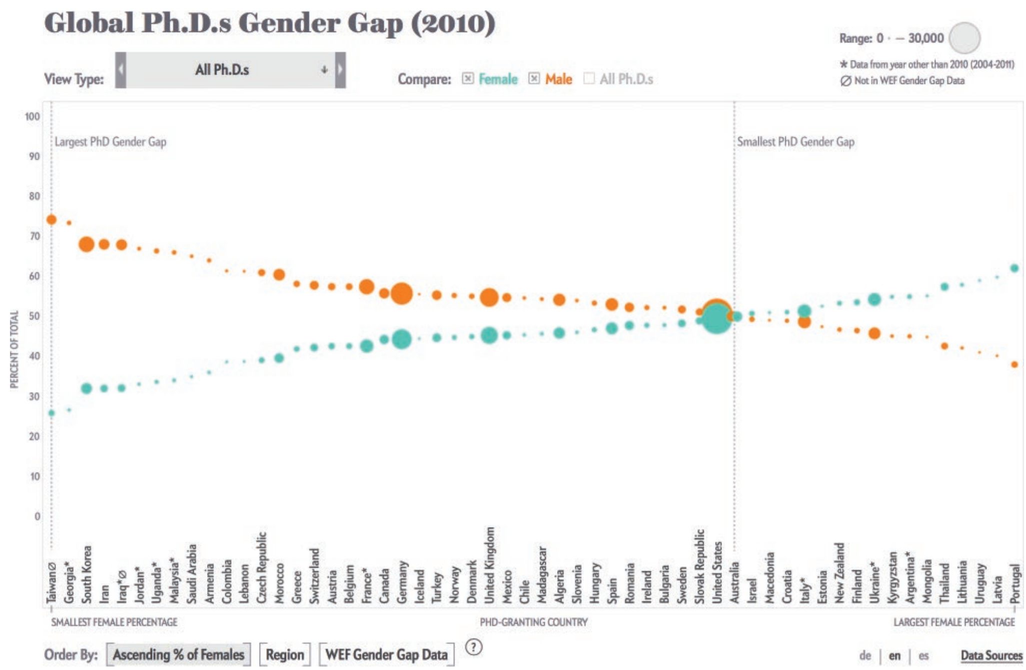


Figure 7.2 How Nations Fare in PhDs by Sex, Interactive by Periscopic; Research by Amanda Hobbs; Published in *Scientific American*

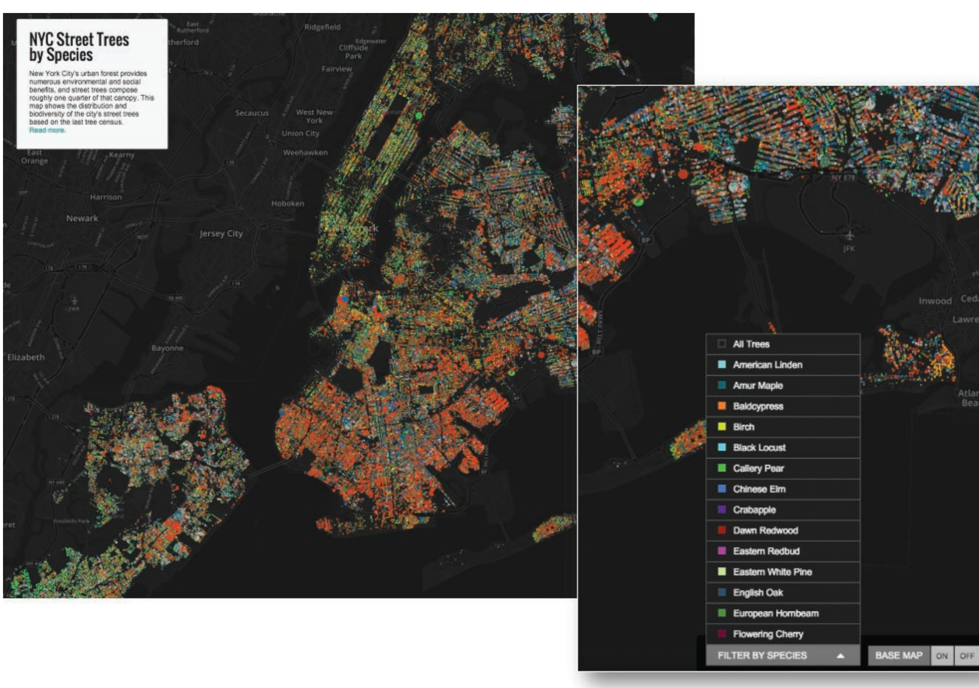


Figure 7.3 NYC Street Trees by Species, by Jill Hubley

To observe patterns for individual tree types is hard: with 52 different tree species there are simply too many classifications to be able to allocate sufficiently unique colours to each, as you will learn about in Chapter 9. To overcome this functionally, the project features a useful pop-up filter list which allows users to adjust the data on view based on revealing only the species of selected interest. Incidentally, notice the big void where JFK Airport is located.

When browsing through the chart gallery you will have already seen the 'treemap' used by the 'FinViz' stock market analysis site (see Figure 6.26). In Figure 7.4 you can see the companion bubble plot used to show this data from a different angle. Here, you can apply a quantitative filter by modifying the positions of the pair of parameter handles along the x- and y-value axes. Doing so will apply changes to the maximum and minimum scales, which means only the values that fall into the new range will be displayed. Further options exist to change the variables plotted on each axis, using the dropdown menu provided.

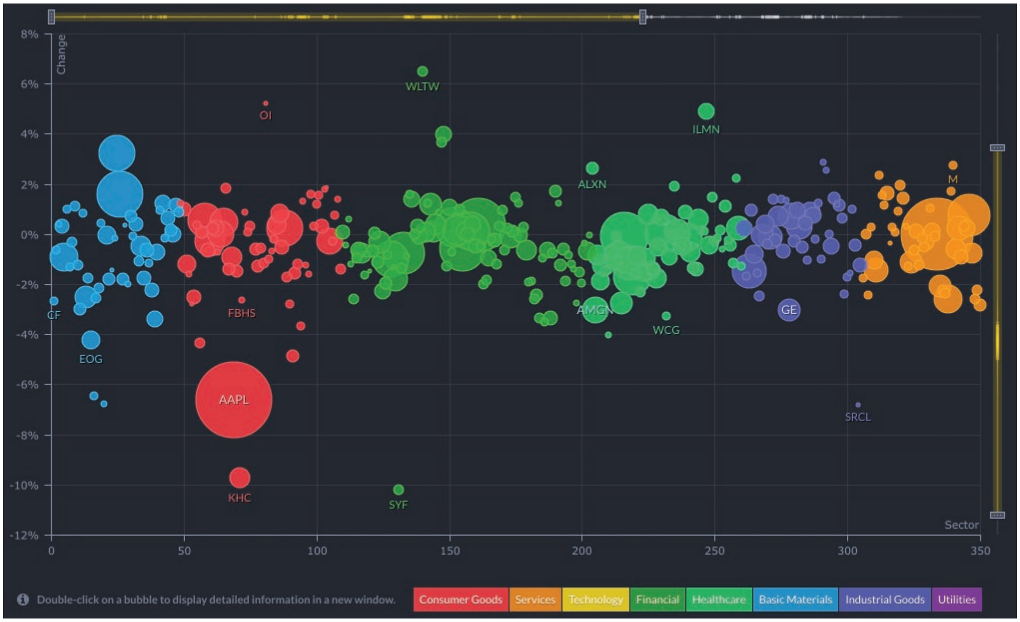


Figure 7.4 Finviz: Standard & Poor's 500 Index Stocks (www.finviz.com)

Highlighting

The second group of interactive features (Table 7.2) offer visual emphasis to highlight data items or values of interest. Whereas the filtering options we have just looked at modified what data items would be included and excluded, these features do not eliminate from a display but create visual or positional contrast. This may be achieved through temporarily modifying attributes of colour or by reordering the arrangement of data items. In contrast to the filtering options, these functions modify the editorial 'focus' perspective.

Table 7.2 Features of Interactivity to Facilitate Highlighting

Example events and controls	Example functions
Select a button or link Select an item from a menu list Select multiple items from a check-box or menu list Select to alter the state of a toggle or radio button Alter the position of a handle along a scale slider Alter the position of two handles along a scale slider (to create a range) Select a mark from within a chart Mouseover a mark from within a chart Select a range of marks from within a chart ('brushing') Type a value into an input box	Highlight selection Highlight values based on selection Highlight associations between selected values Rearrange the order of the data Form calculations based on selection

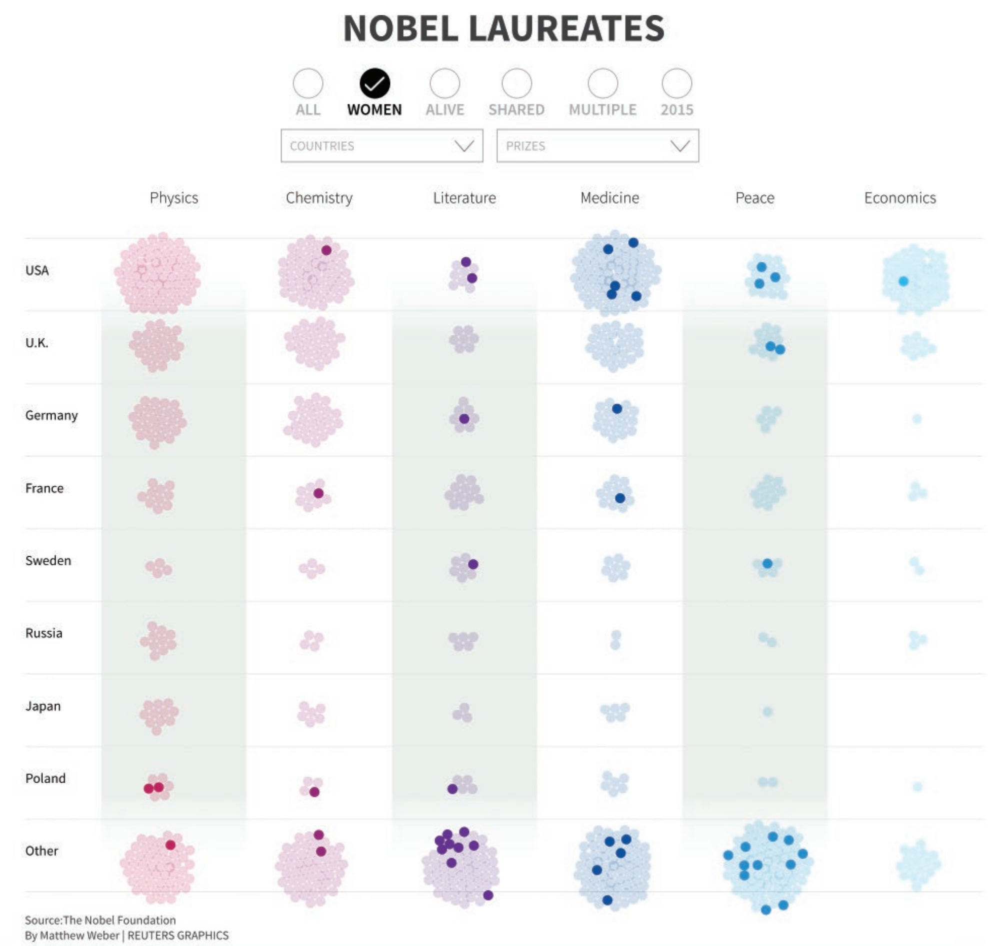


Figure 7.5 Nobel Laureates, by Matthew Weber (Reuters Graphics)

The example in Figure 7.5 demonstrates the use of radio buttons which let you pick different cohorts of all Nobel Laureates to emphasise the matching items across the chart display. As you can see, the selections include focusing on women, shared winners and those who are still living at the time. The selected Laureates are not coloured differently, rather it is the residual values that are significantly lightened to create emphasis through contrast.

In Figure 7.6 we see a bump chart, produced by the Office for National Statistics (ONS), that plots rankings for the 100 most popular names given to baby boys and girls over the past century and beyond. As we learnt in the previous chapter, bump charts can quickly become visually complex when there are multiple items included and the passage of their lines is chaotically up and down. You cannot reasonably colour code 100+ discrete lines for all name categories, so, in this case, users are able to select or enter up to six different names to see how their individual ranking trends have formed over the period.

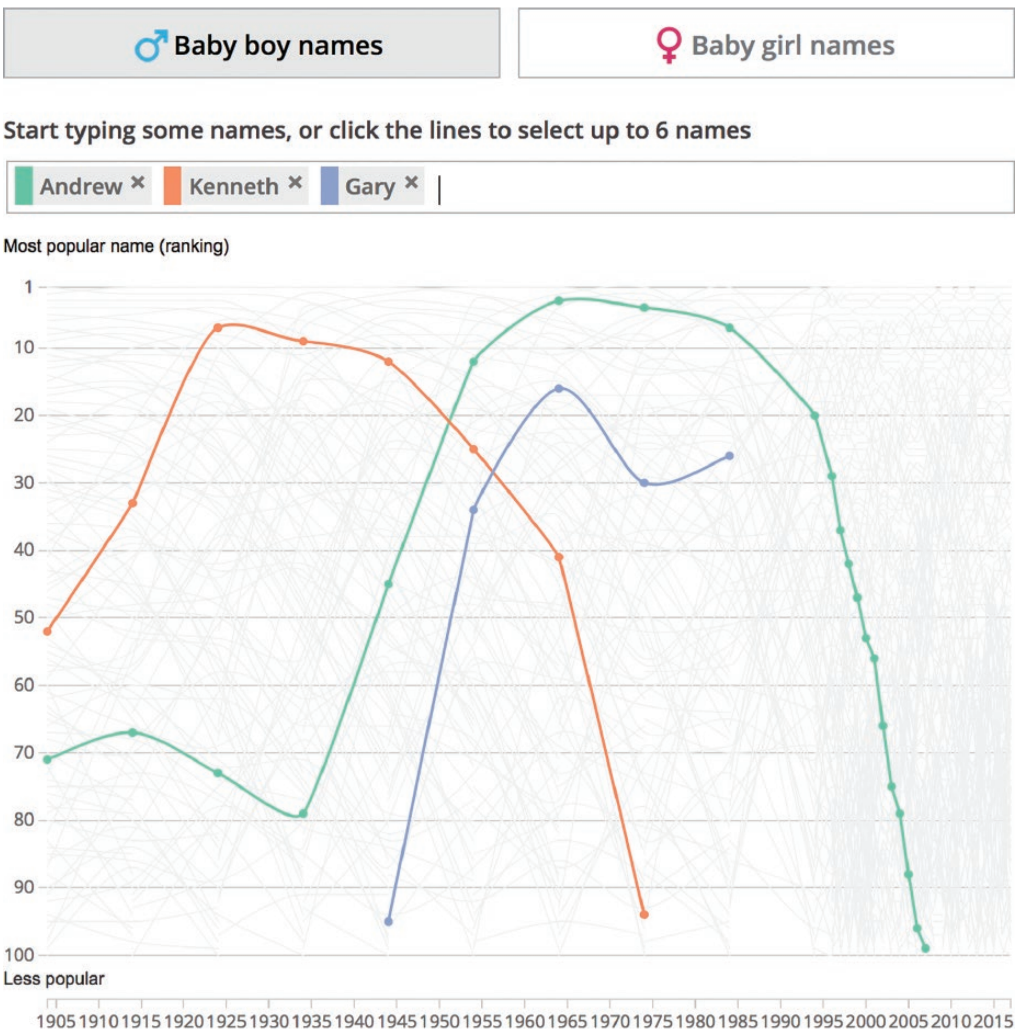


Figure 7.6 Baby Names in England and Wales: 2017, by ONS Digital Content team

The next example (Figure 7.7) portrays the increase or decrease in workers' compensation benefits by US state. This project demonstrates an example of the technique known as data 'linking', where hovering over a mark item in one chart display will highlight an associated item in another chart, thus draw attention to the shared relationship. In this case, hovering over a US state circle, in any of the grid maps, will highlight the same state in the adjacent maps to draw your eye to their respective statuses.

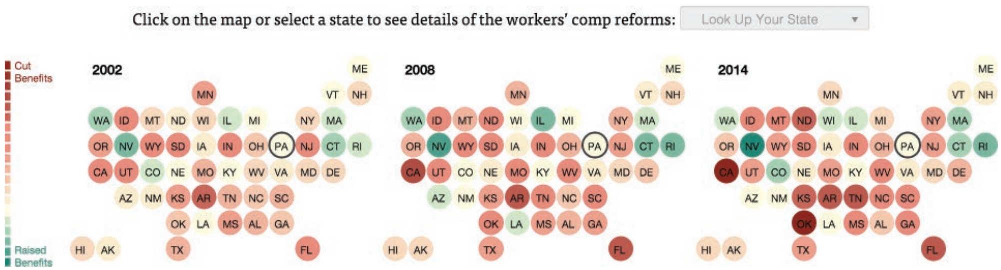


Figure 7.7 Workers' Compensation Reforms by State, by Yue Qiu and Michael Grabell (ProPublica)

Linking and brushing are common approaches used in exploratory data analysis tools where you might have several chart panels and wish to see how selected items compare across each display.

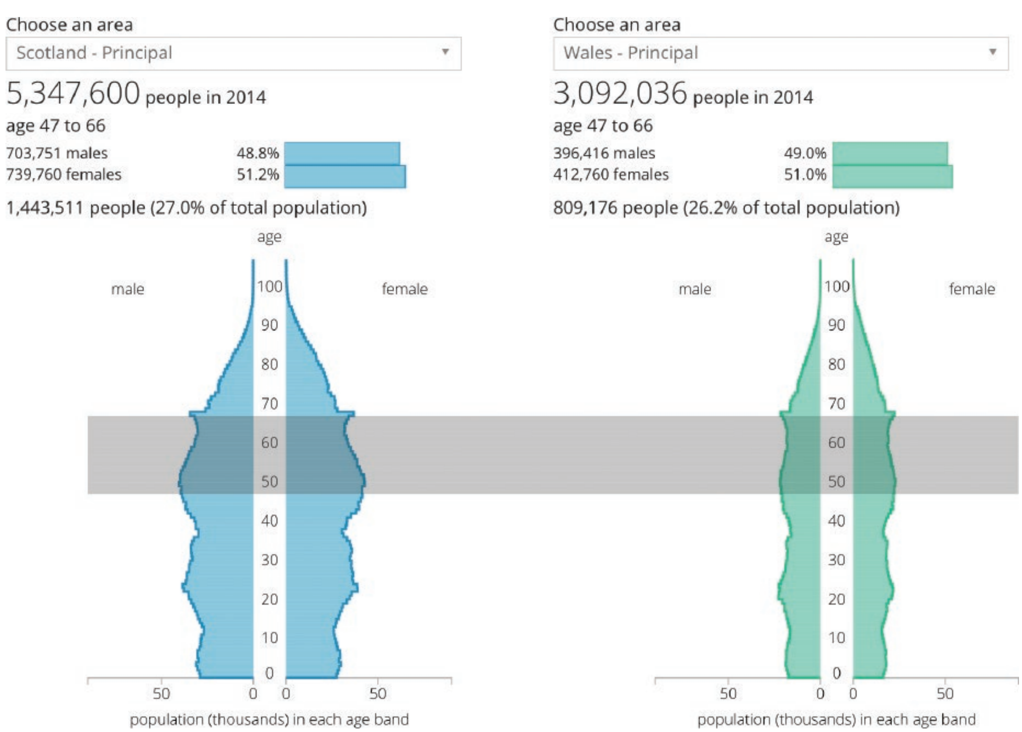


Figure 7.8 How Big Will the UK Population Be in 25 Years' Time?, by ONS Digital Content team

The example in Figure 7.8, once again from the ONS, looks at the UK Census estimates for 2011 and demonstrates both techniques. In this case, you use the cursor to ‘brush’ a selected range of items from within one of the population histograms to inform calculated statistics about the age groups you have chosen, as shown in the panels above the charts. This example also employs ‘linking’ by highlighting the associated regions of the chart in both panels.

In ‘US Gun Deaths’ (Figure 7.9), we see a different approach taken to highlight visually a selection of values of interest. In this piece you can use pop-up check-box lists at the bottom of the page to select different categorical groups. The chosen data items are plotted in the chart view above the baseline separate from the rest, and details of the selection criteria are displayed on the left. Usefully, the ‘remove filters’ option is available in the control panel to reset the display quickly back to the original settings. Note how the transparency of the filter menu allows the data displayed behind it to still be seen. Though it does partially occlude the chart, it is not entirely intruding.

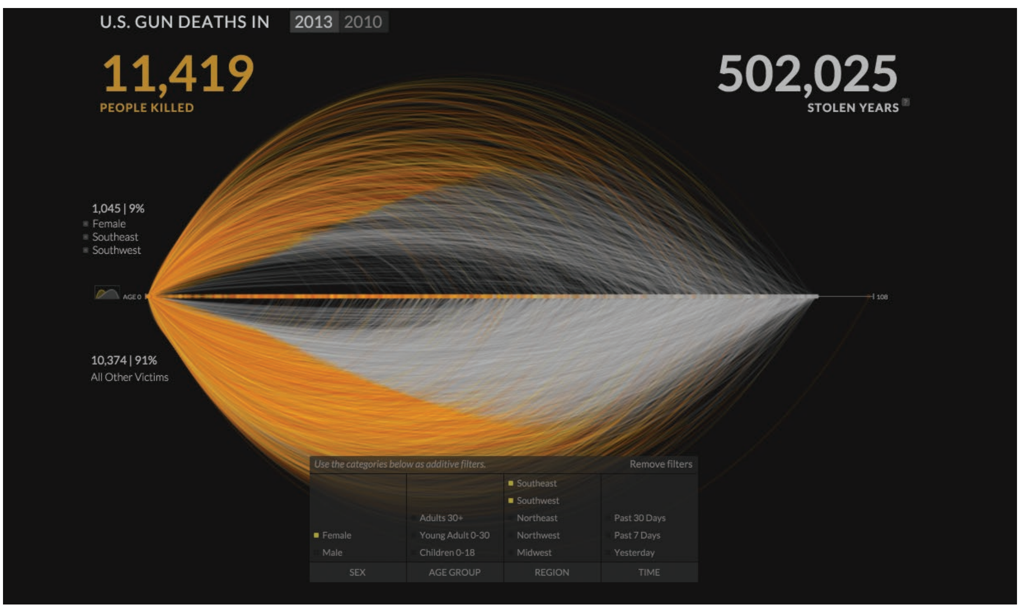
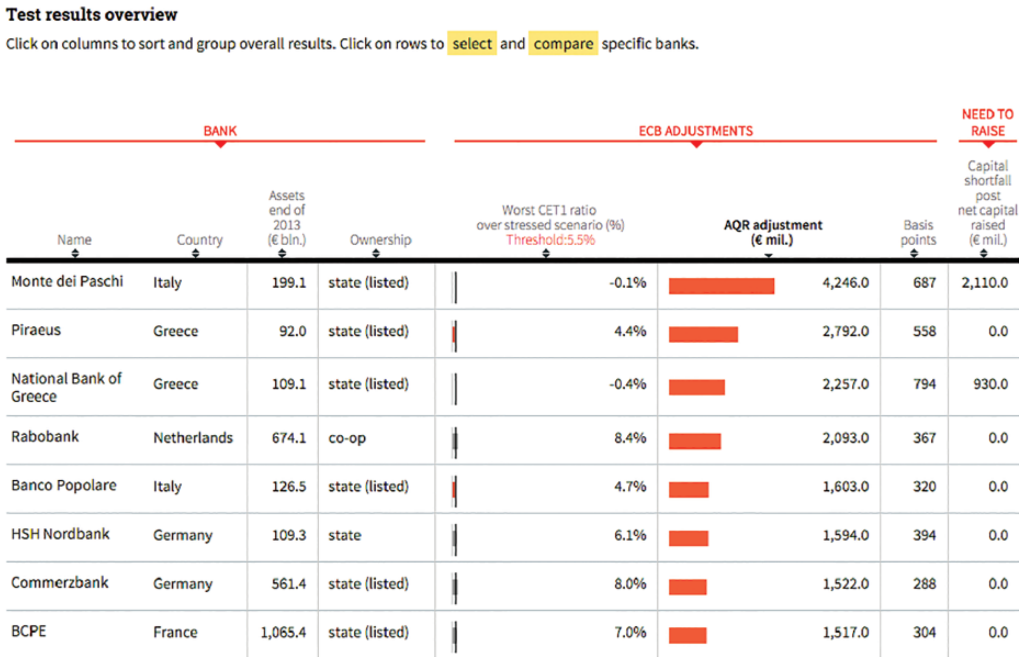


Figure 7.9 US Gun Deaths, by Periscopic

Sorting is another way of highlighting patterns in your data. In Figure 7.10, featuring work by the Thomson Reuters graphics team on ECB bank test results, you can see a tabular display with interactive features in the column headers that allow you to reorder specified columns of data. Columns of categorical data values will be ordered alphabetically; quantitative data values will be reconfigured into ascending or descending order. You can also hand-pick individual records from anywhere in the table to drag them to another position in the table, perhaps to promote them towards the top of the display to facilitate easier comparisons with adjacent records.



Test results overview

Click on columns to sort and group overall results. Click on rows to select and compare specific banks.

BANK				ECB ADJUSTMENTS			NEED TO RAISE
Name	Country	Assets end of 2013 (€ bln.)	Ownership	Worst CET1 ratio over stressed scenario (%) Threshold:5.5%	AQR adjustment (€ mil.)	Basis points	Capital shortfall post net capital raised (€ mil.)
Monte dei Paschi	Italy	199.1	state (listed)	-0.1%	4,246.0	687	2,110.0
Piraeus	Greece	92.0	state (listed)	4.4%	2,792.0	558	0.0
National Bank of Greece	Greece	109.1	state (listed)	-0.4%	2,257.0	794	930.0
Rabobank	Netherlands	674.1	co-op	8.4%	2,093.0	367	0.0
Banco Popolare	Italy	126.5	state (listed)	4.7%	1,603.0	320	0.0
HSH Nordbank	Germany	109.3	state	6.1%	1,594.0	394	0.0
Commerzbank	Germany	561.4	state (listed)	8.0%	1,522.0	288	0.0
BCPE	France	1,065.4	state (listed)	7.0%	1,517.0	304	0.0

Figure 7.10 ECB Bank Test Results, by Monica Ulmanu, Laura Noonan and Vincent Flasseur (Reuters Graphics)

Participating

The techniques presented so far have modified what data you are viewing or how you are viewing it. The next group of features (Table 7.3) involves users taking a more active role by contributing data to help customise a participatory experience.

Table 7.3 Features of Interactivity to Facilitate Participating

Example events and controls	Example functions
Select a button or link Select an item from a menu list Select multiple items from a check-box or menu list Select to alter the state of a toggle or radio button Type values into an input box Alter the position of a handle along a scale slider	Submit data to initiate feedback (e.g. a quiz) Submit data to customise a view

In the project 'Who Old Are You?' (Figure 7.11), users are invited to enter their date of birth using the input box and, based on the calculated age, they are taken to a customised view that compares their age with the ages of a range of famous or celebrated people at the time of major milestone achievements in their lives.

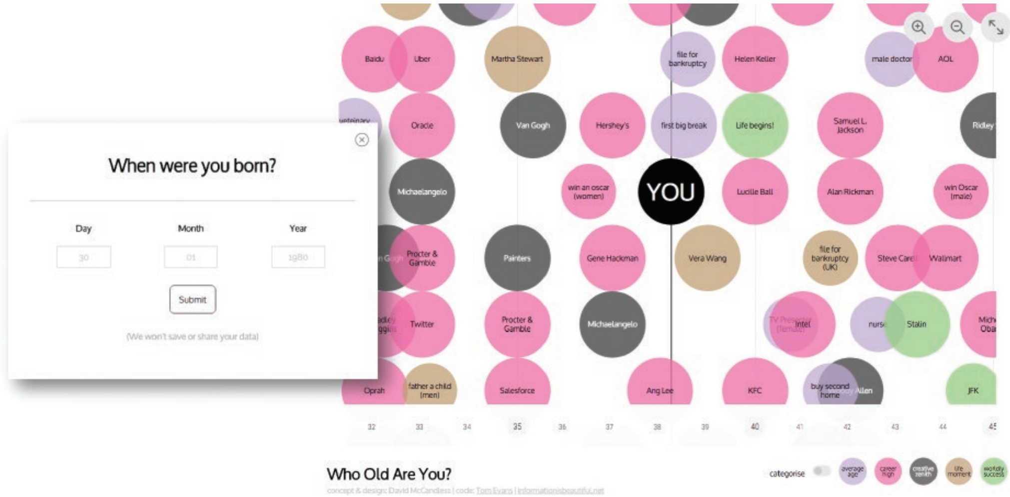


Figure 7.11 Who Old Are You?, by David McCandless and Tom Evans

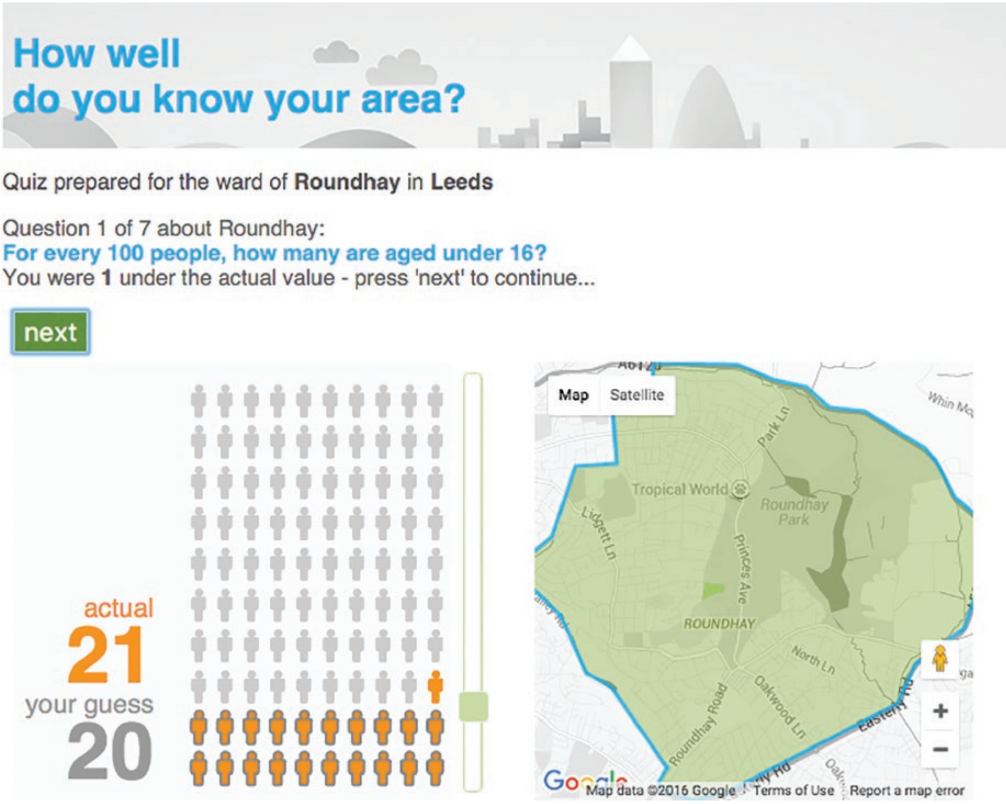


Figure 7.12 How Well Do You Know Your Area?, by ONS Digital Content team

The next example (Figure 7.12), titled 'How well do you know your area?', employs a simple quiz engine to challenge or confirm the knowledge users have about the demographics of their local area. Having entered a UK postcode to establish the neighbourhood, users are asked seven questions, such as 'For every 100 people, how many are aged 65 or over?' To respond, the position of the handle can be modified along the slider to indicate a guess, which will be illustrated by the companion waffle chart. When this estimate is submitted, a correct answer is revealed and an indication of how close or otherwise the guess was, compared with this actual value, is displayed.

Asking people 'what do you think?' and providing immediate feedback to their response is a compelling way to challenge or reinforce people's perceived understanding about

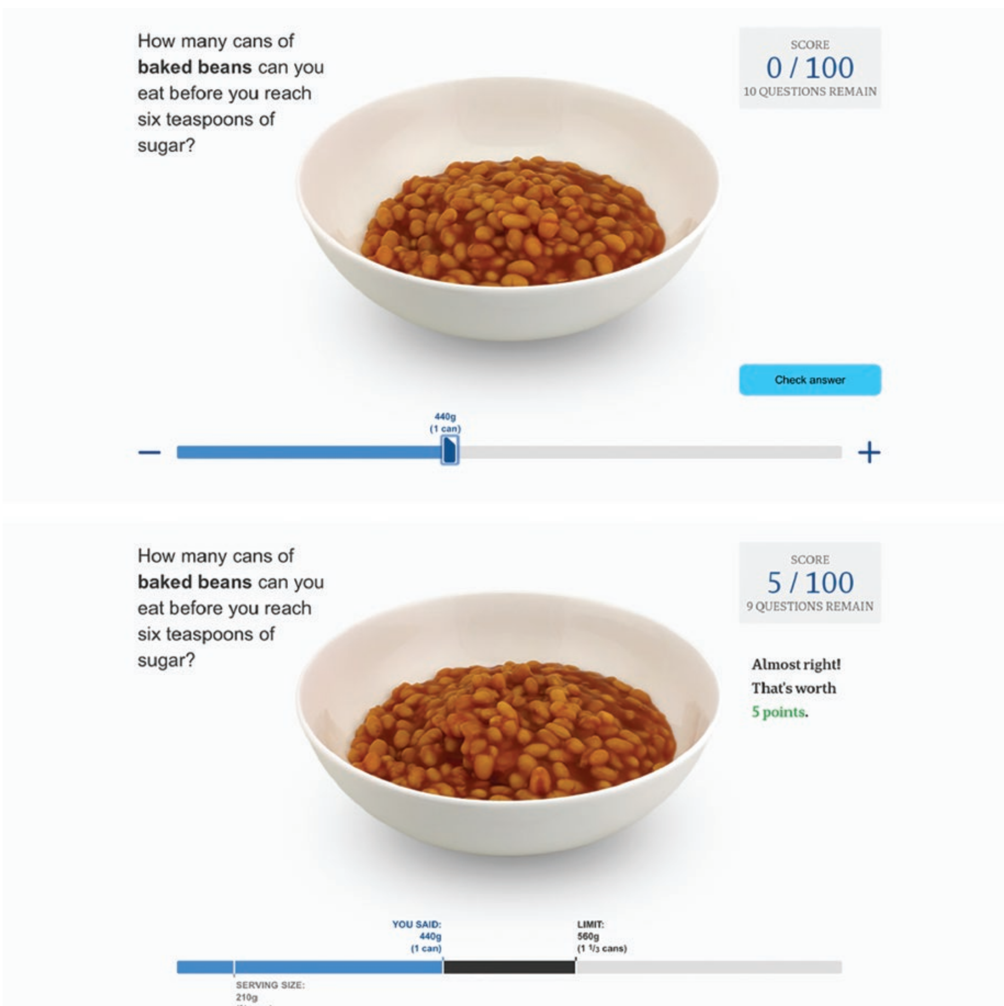


Figure 7.13 Sugar Quiz: How Much Sugar Is in Our Food?, by Claudine Ryan, Ben Spraggon and Colin Gourlay

a subject. In this next work by ABC in Australia (Figure 7.13), a similar approach is taken to ask people 'how much sugar is in our food?' using a series of 10 questions to test participants' knowledge of the sugar content in some of the most popular groceries. Each question is framed the same way, asking how much of a given item can be consumed until six teaspoons of sugar have been reached. Like the ONS quiz, users enter their estimates using a slider, with a nice additional feature being the modified imagery to offer a visual that matches your estimate.

The next example is a project that records and reuses the data it collects. Figure 7.14 shows screens from 'Do you remember where Germany was divided?', where users are invited to draw a line representing their estimate of the route of the former border between East and West Germany. There is no assistance provided in terms of town or city markers, so, using the pencil cursor, you draw a line to reflect your best recollection of its shape. When you have completed your drawing, a more detailed map view is automatically loaded up to place your suggested route in the context of the actual route. Additionally, having collected and saved the drawings of other participants, it shows and calculates a comparison of how accurately your drawing was compared to other people.



Figure 7.14 Do You Remember Where Germany Was Divided? [Translated], by Berliner Morgenpost/ Funke Interaktiv

Annotating

As mentioned numerous times, some ways of representing data do not place an emphasis on users being able to judge values easily to any degree of precision. This might be entirely consistent with the intended tone of a project. When interactivity is available, it is possible to address the quite reasonable appetite some users may have to see more details about the data they are seeing.

Providing a temporary display of data details ‘on demand’ can be achieved in different ways (Table 7.4). One issue to be aware of when creating pop-up tooltips is to ensure the place in which they appear does not risk obstructing the view of other adjacent marks in the region you are currently interrogating. This can be an intricate thing to handle, especially when you have a lot of annotated detail to share.

Table 7.4 Features of Interactivity to Facilitate Annotating

Example events and controls	Example functions
Select a link or button Select a mark from within a chart Mouseover a mark from within a chart	Reveal annotations in a local tooltip/pop-up Reveal annotations in a separate panel

The example shown in Figure 7.15 uses a heatmap to show the relationships between different ‘Avengers’ characters. Specifically, it plots how often the main characters have appeared together in the same comic book titles over time. The colour coding applied to a heatmap lets users form a general sense of the main patterns of frequent and infrequent shared appearances. For those who want more detailed values, however, they can simply hover over the chart and click on the intersecting cell of interest. The space available to the right of the chart is then occupied with a detailed annotation presenting images of the pair

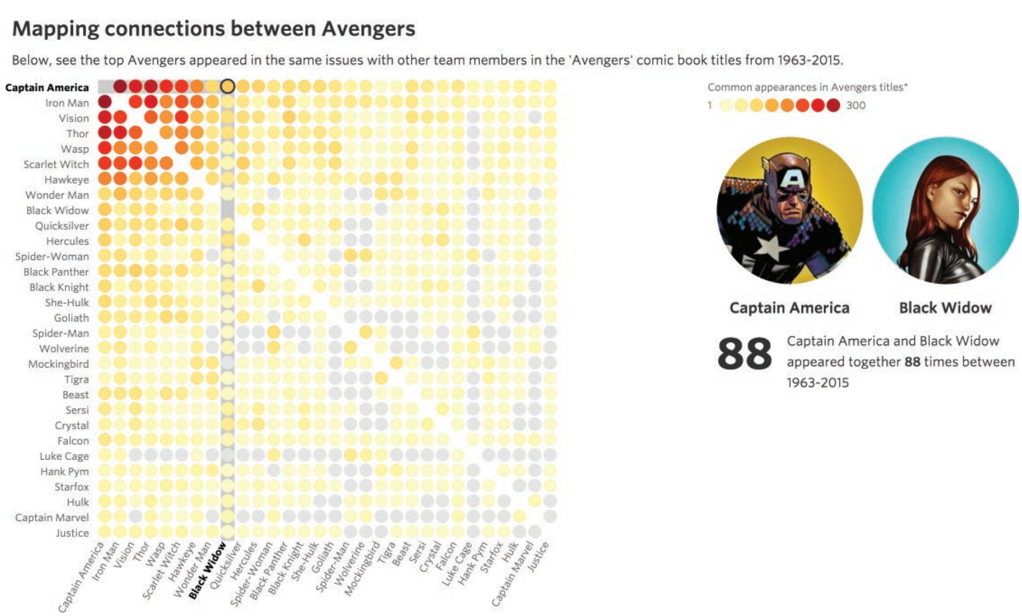


Figure 7.15 How the ‘Avengers’ Line-up Has Changed over the Years, by Jon Keegan (*Wall Street Journal*)

of chosen characters as well as the statistic for how often they have appeared together. Notice also how background shading and bold font labels are added to the heatmap to help orientate which row and column has been selected.

Not all project layouts provide sufficient empty space to accommodate annotation in this way and so pop-up displays offer a solution to overcome spatial constraints. The example in Figure 7.16 analyses the rhetoric used by US Presidents through history in the annual 'State of the Union' address. Circle marks are proportionally sized to indicate the standardised frequency of different words being used in the speeches given by each president over time. This display facilitates a sense of the main patterns, but, to learn about the exact values, users can hover over each circle to bring up an annotated tooltip displaying details of which year, which president and how frequently the given word or phrase of interest was used.

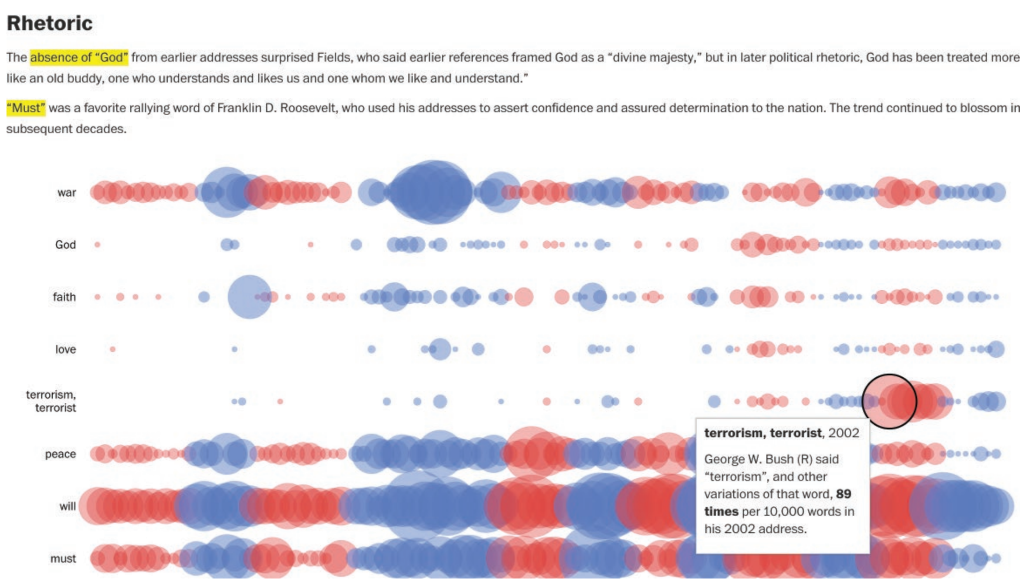


Figure 7.16 History Through the President's Words, by Kennedy Elliott, Ted Mellnik and Richard Johnson (*Washington Post*)

As you will learn in the next chapter, annotation is about providing useful assistance to your users. One key potential feature of assistance can be a 'how to read' guide, helping users to understand how to read chart types.

Connected scatter plots are unfamiliar chart types to many audiences. Recognising that users may not necessarily understand how to read them, Bloomberg's visual data team provide a pop-up 'How to read this graphic' guide when they visit the project shown in Figure 7.17. This guide can be closed but remains available for those who may need to refer to it again. The connected scatter plot was the right choice for this analysis, showing the relationship between

two quantitative variables over time. Rather than use a different and possible inferior representation approach, it is to the authors' credit that they respected the capacity of their users to learn how to read this graphical form.

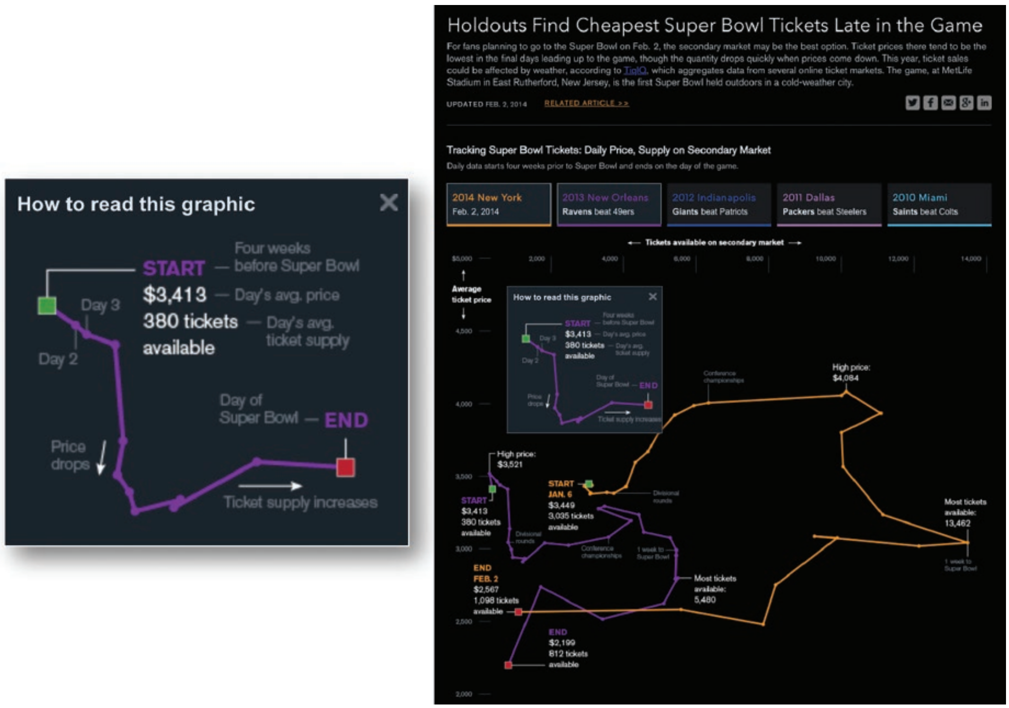


Figure 7.17 Holdouts Find Cheapest Super Bowl Tickets Late in the Game, by Alex Tribou, David Ingold and Jeremy Diamond (Bloomberg Visual Data)

Animating

Data that has a temporal dimension can present opportunities for being displayed using some form of animated sequencing. Seeing values transition over time can sometimes expose interesting patterns that may otherwise be hidden or imperceptible through comparing static views in isolation.

Table 7.5 Features of Interactivity to Facilitate Animating

Example events and controls	Example functions
Load a web page Select a button (play, pause, stop and reset, speed buttons) Alter the position of a single handle along a scale slider	Automatically initiated animation Manually initiated animation (using buttons) Manually controlled animation (using a slider)

In some respects, using the label of interactivity to describe the functions (Table 7.5) of an animated visualisation can be misplaced. For some animated pieces, one could argue they are more a matter related to *composition* thinking and indeed, oftentimes, they will not actually be controllable by any means. The increasingly popular animated gif is such an example, whereby you effectively open it and it runs automatically.

In many cases there will be at least some control for starting and stopping an animated sequence, like this first example in Figure 7.18, titled ‘Breathing Earth’. This work simulates the health of vegetation around the planet between states of lush and arid, pulsing in different ways in different places through the seasons of a year. The more a region is covered in a darker colour shade indicates a greater measure of ‘greenness’, the shorthand term used to represent the scientific vegetation index. The main purpose of this piece is to witness the data presented in this dynamic fashion and to experience repeatedly this animated loop in order to find new seasonal and spatial observations. As expressed by Nadieh Bremer in the description that accompanies her work, ‘the more often you watch the year go by, the more the small details will start to stand out’.

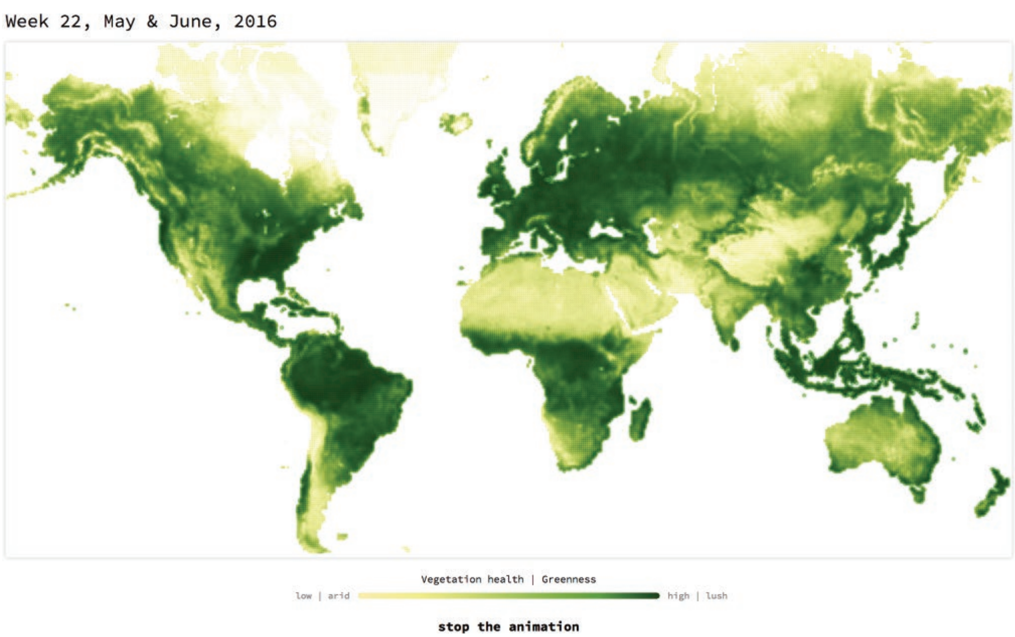


Figure 7.18 Breathing Earth, by Nadieh Bremer (Visual Cinnamon)

The next example (Figure 7.19) plots the distribution of height and weight of NFL footballers over time and presents the data using an animated heatmap to help reveal this shifting pattern. When you land on the page, the animation automatically initiates, but you can then assume control by using the play button to start and stop the animation when you wish. Additionally, you can grab and move the handle along the time slider manually to reposition the time frame view.

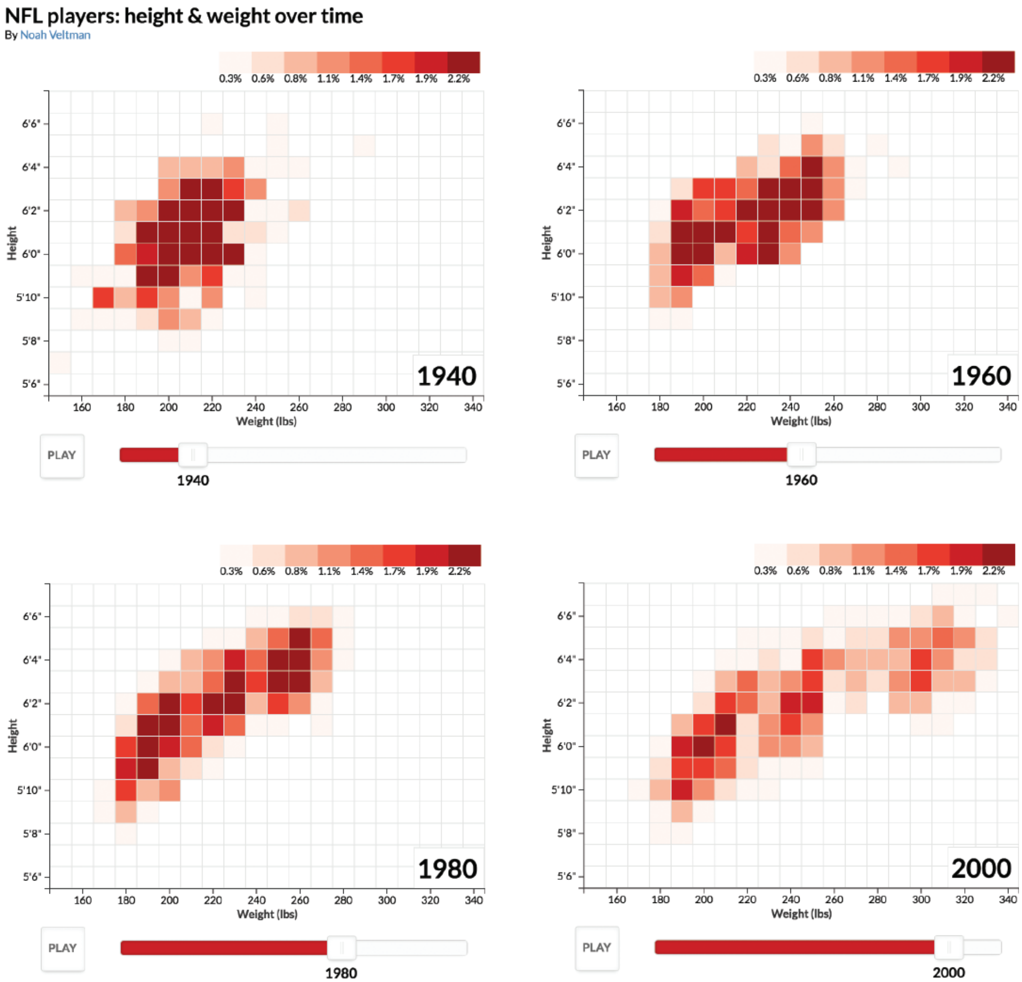


Figure 7.19 NFL Players: Height and Weight over Time, by Noah Veltman (noahveltman.com)

Navigating

One of the main benefits of interactivity is to overcome the limitations of space. You might have lots of detail or contents to share but not enough room to make it reasonably and simultaneously accessible. The next group of dynamic features (Table 7.6) enable users to access multiple views or explore greater levels of detail.

Table 7.6 Features of Interactivity to Facilitate Navigating

Example events and controls	Example functions
Select a button (such as zoom level)	Zoom in and out of a scaled level of detail
Select tab elements (such as a dot stepper)	Navigate ('pan') around a detailed display
Scroll in or out	Navigate through a sequence of discrete pages

Example events and controls	Example functions
Select a region from a map or menu Select, hold and draw a region of interest Select, hold and move Alter the position of a single handle along a scale slider Sideward scroll (unique to trackpads, Mac Mouse)	Navigate through a sequence of displays (within the page) Navigate through a gradual unveiling of a visualisation

The first example is simply titled 'Earth' (Figure 7.20) and offers a powerful, elegant and widely used tool to explore live patterns of wind, weather and ocean conditions anywhere on the planet. It offers a multitude of different interactive features, including another demonstration of data being animated – wind, weather and the oceans are clearly phenomena that lend themselves to dynamic representation. However, for the scope of this section of features, it is the enabling of users to navigate and view the map across any position around the globe – known as 'panning' – and to adjust the scale of their view – known as 'zooming'.

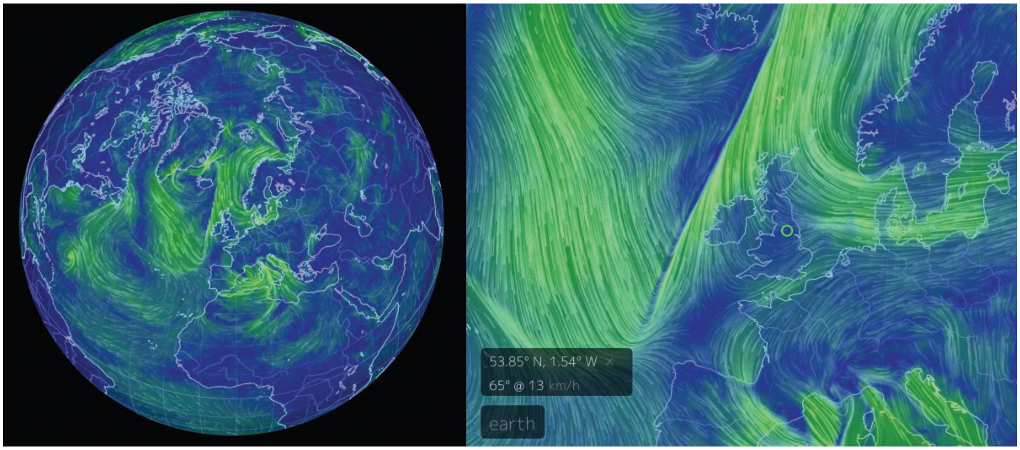


Figure 7.20 Earth, by Cameron Beccario (earth.nullschool.net)

The dot map in Figure 7.21 displays an incredibly detailed representation of population density across part of the USA. There are over 300 million dots plotted, one for each person residing in the USA, colour coded by race and ethnicity, based on data from the 2010 Census. Like the Earth wind map, users can pan around different locations on the map and then use a scrollable zoom or scaled zoom buttons incrementally to change the level of detail displayed. This act of zooming to increase the magnification of the view is known as a geometric zoom, effectively re-framing the included and excluded data through the window at each scale level.

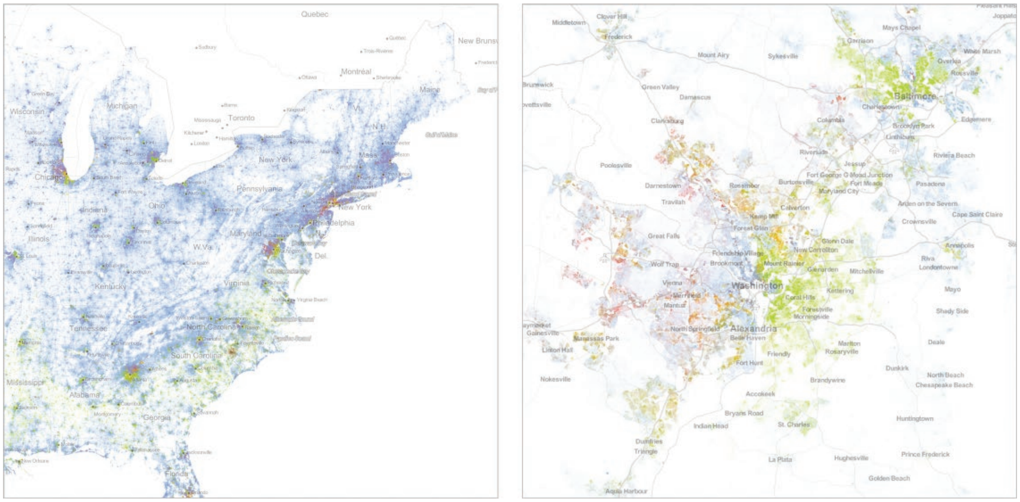


Figure 7.21 The Racial Dot Map: Image Copyright, 2013, Weldon Cooper Center for Public Service, Rector and Visitors of the University of Virginia (Dustin A. Cable, creator)

Both these recent works demonstrate methods for navigating vast landscapes in detail, enabling users to dictate which views and levels of scale to explore the data through. An alternative approach to navigating involves the offer of a more linear, explanatory experience,



Figure 7.22 Killing the Colorado: Explore the Robot River, by Abrahm Lustgarten, Al Shaw, Jeff Larson, Amanda Zamora and Lauren Kirchner (ProPublica) and John Grimwade

building up a narrative about a subject through a series of discrete sequences. It is arguably the quintessential example of storytelling with data and is often presented using a technique known as 'scrollytelling', whereby users scroll to move vertically up and down through the steps of a story.

The project featured in Figure 7.22 is a prime exhibit of this kind of dynamic interface. It offers a step-by-step journey down the length of the Colorado River to investigate the impact of some of the major infrastructure projects that have caused the gradual draining of this vital source of water for millions of Americans. To break out of the linear navigation, users are also able to jump ahead or back to different chapters of interest using the left-hand menu. This can be a particularly useful feature after you have been through the full sequence once. Another helpful device is the inclusion of a thumbnail image to help orientate the location of current focus within the context of the overall journey down the river.

Sequentially building up a story can prove to be a powerful way of facilitating understanding. In Figure 7.23, the project 'What's Really Warming the World' presents a sequence of possible causes for climate change. As you scroll down the page (or, alternatively, click through the page steppers or directional arrow) it takes you through the different hypotheses, overlaying data about each onto a chart plotting the observed changes in temperature. Eventually, you reach the conclusion and the big reveal: it is due to greenhouse gases.

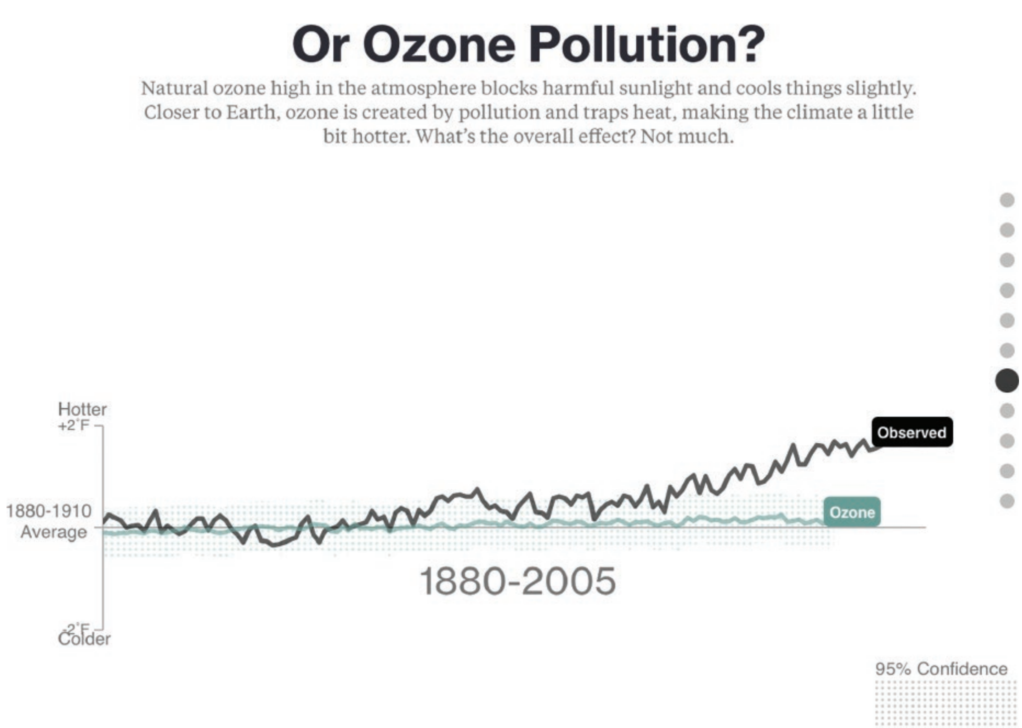


Figure 7.23 What's Really Warming the World?, by Eric Roston and Blacki Migliozzi (Bloomberg Visual Data)

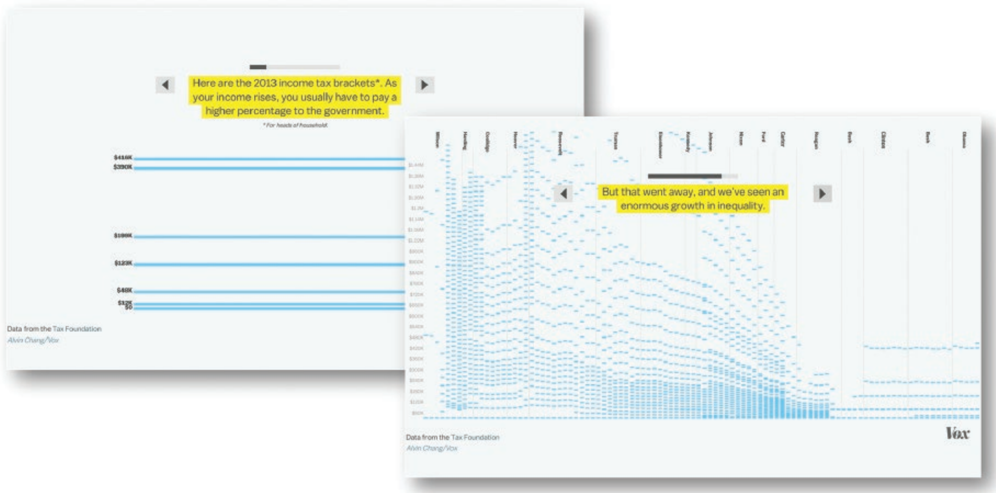


Figure 7.24 100 Years of Tax Brackets, in One Chart, by Alvin Chang

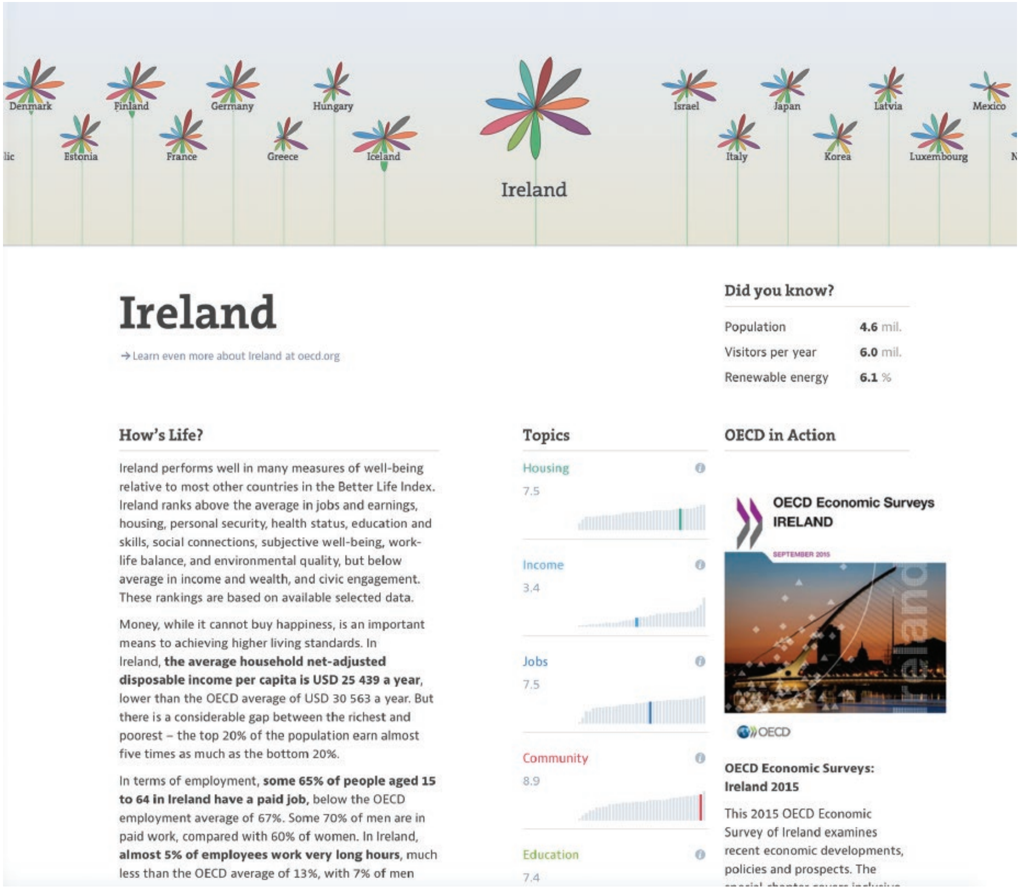


Figure 7.25 OECD Better Life Index, by Moritz Stefaner and Dominikus Baur, Raureif GmbH

The work shown in Figure 7.24 looks at 100 years of tax brackets in the USA. This project employs a similar sequencing approach to unfold content, but rather than stepping through a series of different charts or discrete views of a subject, the sequence here gradually builds up the user's understanding about the subject matter. It steps through information about why tax brackets are a relevant topic, what they are, how they affect people, and what are some of the main historical patterns to have emerged from this analysis. When you have technical topics like this it might trigger indifference among an audience through their lack of domain knowledge. So, rather than drop them straight into the deep end, a skilfully executed and carefully considered step-by-step presentation like this, coaching them rather than exposing them, can lead to increased engagement.

A final demonstration of a navigating device is characteristic of a 'drill-down' feature, giving users access to data which might exist at a lower granularity or hierarchical level. The flower representations used in the Better Life Index project (Figure 7.25) effectively co-exist as menu items. These offer ways of choosing a country to navigate to a separate report that provides far more detailed analysis and commentary, supplementing the summarising data as displayed in the initial flower view.

7.2 Influencing Factors and Considerations

You should now have established a good sense of the wide range of possibilities for incorporating interactive features into your work. So let's turn our attention to consider the factors that will have most influence on which of these techniques you might need to or choose to apply.

Constraints: The main factor that will shape the scope for employing interactivity is unquestionably the technical skills you possess and the capabilities of the technology you have access to. If you are technically limited in being able to develop any of the features profiled, it immediately rules them out of your thinking. In the online resources that accompany this book, I include a guide through some of the contemporary applications, tools and programming libraries that enable you to develop visualisations with interactive features. So, looking beyond technology for now, another major constraint will exist through the pressure of time. If you have a limited time frame in which to complete your work, you are going to find it a challenge to pursue particularly ambitious or bespoke interactive solutions, even if you possess extensive technical competence.

Deliverables: Understanding the expected deliverables of your work is vital. Just because a visualisation might be created and published digitally, the output may still be non-interactive: *digital* does not necessarily or automatically mean *interactive*. If your work is for print only, interactivity is not needed, and this entire chapter of thinking will be outside your radar of concern.

The main question to ask is whether the characteristics of the setting in which your audience will consume a visualisation are compatible with the prospect of needing them to interact. Will your audience have the time, the patience and indeed the know-how to exploit such features?

Additionally, what are the varied device specifications on which your solution will need to function? To what extent will you be seeking to emulate the same experience across mobile, tablet and desktop devices? How adaptable might your solution need to be, and might there be a need for compromises?

Where once we were limited to the mouse or the trackpad as the common peripheral, over the past decade we have seen the emergence of touch-screens, through smartphones and tablets. This has introduced a whole new challenge for developers to find ways of creating consistent solutions that are compatible across different platforms.

For simplicity and consistency, in this chapter we have focused on events associated with using a mouse or trackpad, but here is a translation of the equivalent touch events (Table 7.7). The primary difference between the two concerns the inability to enact the equivalent of a hover (or 'mouseover') event with touch-screens.

Table 7.7 Comparison of Interaction Event Types

Mouse/trackpad event	Touch event
Left click, right click	Single-finger tap, two-fingered tap
Double click	Double tap
Click, drag and drop	Tap, drag and drop
'Mouseover' or pointer 'hover'	Tap
Wheel scroll	Swipe (move), pinch/reverse pinch (for zoom)
Unique: keyboard controls	Unique: rotate

Purpose: Though all visualisations that offer exploratory experiences will need to be interactive, not all interactives are exclusively for offering exploratory experiences. Some of features we have profiled, such as those that navigate through sequences, are classic demonstrations of using interactivity to fulfil an explanatory experience.

Your definition about what experience to offer will inform the features of interactivity you may seek to incorporate. As mentioned in Chapter 3, there may also often be scope for a blended approach. For instance, you might open a visualisation with an explanatory experience, based around showing some main findings and telling your audience something. Through interactivity, you may then transition the users towards more of an exploratory interface that invites them to interrogate the data to pursue their own particular curiosities about the subject.

Data representation: Some charts are inherently visually complex, and this can create obstacles for the user trying to understand them. The bump chart, chord diagram and Sankey diagram are just a few of the charts that commonly have multiple lines and bandings crossing over in the same space. Offering interactive features that enable filtering and/or highlighting of certain data items can help them become a more palatable prospect for the user to engage with. The Sankey diagram shown in Figure 7.26 looks at a month of data about animals entering a shelter. The left

side of the diagram displays the origin stories which are connected with the proportional bands to the associated outcomes on the right side. With so many crossing paths, the ability to choose a single category on the origin or outcome side of this display helps reveal some of the important stories more discernibly. The chart becomes more readable and, by extension, more usable.

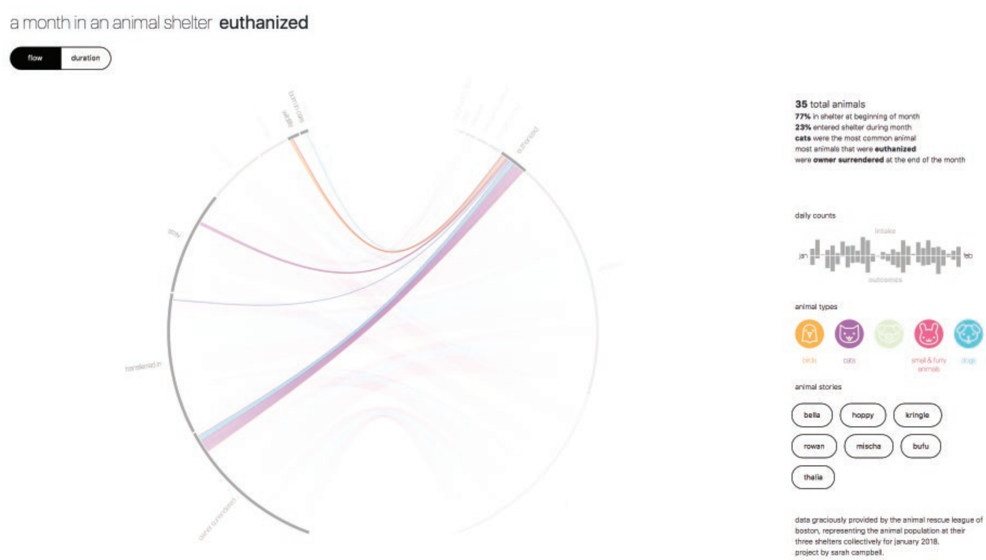


Figure 7.26 A Month in an Animal Shelter, by Sarah Campbell

Trustworthy design: The reliability, consistency and functional performance of a visualisation is something that influences the perceived 'trustworthiness'. Does it do what it promises, and can the user trust the functions that it performs?

> 'Confusing widgets, complex dialog boxes, hidden operations, incomprehensible displays, or slow response times … may curtail thorough deliberation and introduce errors.' **Jeff Heer and Ben Shneiderman, taken from 'Interactive Dynamics for Visual Analysis' (2012)**

Inevitably, issues around data privacy and intended usage of data collected will be important matters to handle with integrity and transparency, otherwise the trustworthiness of your work may be critically undermined. In most projects, any data contributed by a user is collected only for a temporary period of participation (i.e. not held beyond the moment of usage). However, if you intend to collect and save this data, perhaps to append to an original dataset and use this to improve the content, you should make your intentions very clear to the user.

Is it a one-off piece of work or something that will run on regularly updated data? In which case how robust is the design going to be to accommodate new data? Will it cope with new categories and larger or smaller value ranges? Who will keep running it and who will support it thereafter to ensure it continues to offer a quality experience that preserves the trust of its users?

Accessible design: Seek to minimise the friction between the act of using an interactive feature and the understanding it facilitates. Do not create unnecessary obstacles that stifle curiosity. Indeed, resort to interactivity only when you have exhausted the possibilities of an appropriate and effective static solution.

For example, let's consider features of animating. If your data is not changing much, an animated sequence may be of no merit. If it is changing a lot, maybe it will be too chaotic and the movements too sudden to be observable. Your intention may have been to exhibit this chaos, but the main value of animated sequences should be to help reveal dynamic patterns of change rather than random variation.

It really depends on what it is you want to show: the dynamics of a 'system' that changes over time or a comparison between different states over time? With animated sequences, there is a reliance on memory to conduct a constant comparison of change. Yet, our recall ability is fleeting at best and weakens the further apart (in time) the basis of the comparison has occurred.

The speed of an animation is also a delicate matter to judge as you seek to avoid the phenomenon of change blindness. Rapid sequences will cause the stimulus of change to be missed; a tedious pace will dampen the stimulus of change and key observations may be lost. Access to the available understanding becomes diminished.

> 'Generations of masterpieces portray the legs of galloping horses incorrectly. Before stop-gap photography, the complex interaction of horses' legs simply happened too fast to be accurately apprehended … but in order to see the complex interaction of moving parts, you need the motion.' [Paraphrasing] **Barbara Tversky and Julie Bauer Morrison, taken from *Animation: Can it Facilitate?***

If you wish to facilitate direct comparisons you ideally need to juxtapose individual frames within the same view. The common technique used to achieve this in visualisation is through *small multiples*, where you repeat the same representation for each moment in time of interest and present them collectively in an adjacent view, often through a grid layout. This enables far more incisive comparisons. The famous 'Horse in Motion' work by Eadweard Muybridge (Figure 7.27) was carried out to learn about the galloping form of a horse by seeing each stage of the motion through individually framed moments.

Elegant design: The self-discipline required to avoid the temptation of feature creep is indisputable at this stage of the design process. For those people with a natural technical flair, there can be a strong temptation to incorporate interactivity when it is neither required nor helpful: just because you *can* does not mean to say you *should*.

Judging the degree of flexibility is something of a balancing act: you do not want to overwhelm the users with more adjustments than they need, nor do you want to narrow the scope of their likely interrogations. If the characteristics of your audience are varied, you may be understandably inclined to try including more features than are necessary. For a one-off project you have to rely on your own best judgement; for projects that will be repeatedly used you will have more potential to seek and accommodate feedback to inform refinements.

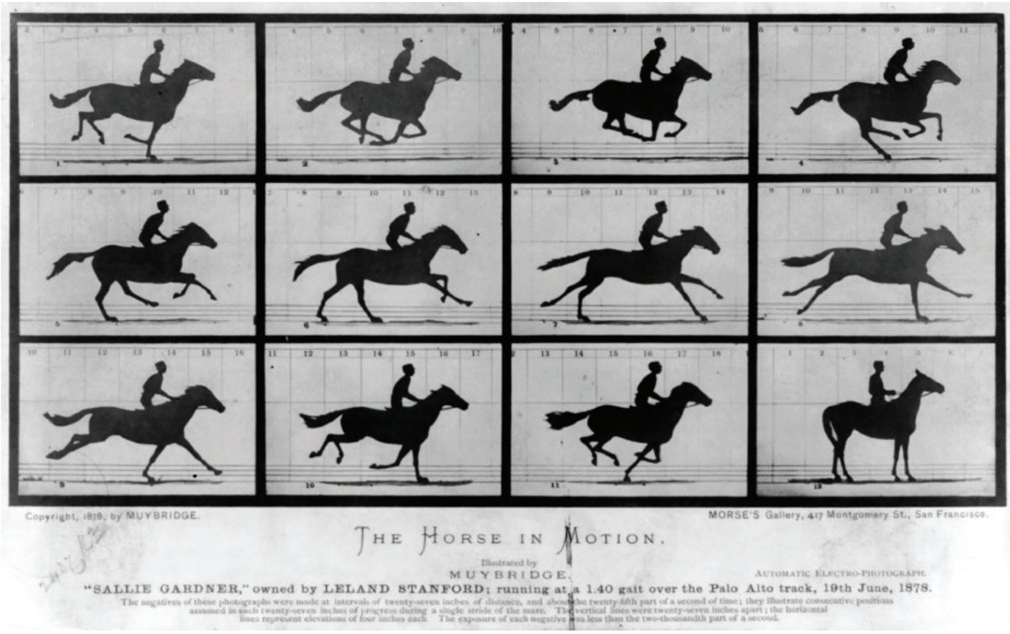


Figure 7.27 The Horse in Motion, by Eadweard Muybridge.

Source: United States Library of Congress's Prints and Photographs division, digital ID cph.3a45870

Elegance in designing interactive features extends to their appearance and the seamlessness with which they can be accessed and used. With visualisation you are aiming to make invisible insights visible. Conversely, the features of visible design should be as inconspicuous and intuitively packaged as possible.

The captivating quality of a well-conceived interactive visualisation is how it can introduce new means of engaging with data that simply could not have been delivered without the incorporated features. It is important to acknowledge that there can be pleasure created by thoughtfully conceived interactivity. Even if there are features that offer only ornamental benefit, a sense of fun and playability can be appealing to any audience type, so long as the circumstances are right and such features do not obstruct access to understanding.

Summary: Interactivity

Features of Interactivity

This chapter introduced the potential value of incorporating interactive features into your work, profiling a wide range of options that will enable users to interrogate and control a visualisation. These included:

- Filtering: Enabling users to specify what data they wish to include or exclude from a chart display.
- Highlighting: Features that apply visual emphasis to highlight data items or values of interest.

- Participating: Inviting users to contribute data to help customise a participatory experience.
- Annotating: Offering users ways to see more details about the data they are seeing.
- Animating: Displaying data with a temporal dimension using animated sequencing.
- Navigating: Features that enable users to access multiple views or explore greater levels of detail.

Influencing Factors and Considerations

If they were the options, how did you make your choices? The influencing factors included:

- Constraints: The technology and skills possessed, as well as the timescales, will shape ambitions.
- Purpose: What experience are you facilitating and how might interactive options help achieve this?
- Data representation: Certain chart choices may require interactivity to enhance readability.
- Trustworthy design: Functional performance and reliability will substantiate the perception of trust from your users.
- Accessible design: Interactive features should be useful and unobtrusive – minimise the clicks.
- Elegant design: Beware of feature creep but embrace the potential of fun and playability.

General Tips and Tactics

- Good project management is critical when considering the development of an interactive solution.
- Do not be distracted by working on interaction features that seem 'cool' or 'fancy', but do not add enough value to warrant precious resources being allocated (time, effort, people).
- Keep focusing on what is important and relevant. A technical achievement may be great for you and your CV, but is it needed for the project?

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

8

Annotation

The third element of the visualisation design anatomy is annotation. This concerns judging the level of assistance an audience may require in order to understand the background, function and purpose of a project, as well as what guidance needs to be provided to help viewers perceive and interpret the data representations.

In contrast to the more theoretical and technical concerns around data representation, colour and interactivity, judgements about what annotated features to offer your viewers can be more heavily informed by common sense. This is an influential but often neglected layer of thinking that really exposes the amount of care a visualiser shows towards the audience.

The sequence of suggestions roughly follows the typical organisation of the layout of your work, beginning with the features that might typically exist at the start of an experience, working through to those usually found towards the end. Towards the end of the chapter we will look at the factors that will influence your choices, but first let's profile some of the key features of annotated design you might consider including in your visualisation.

8.1 Features of Annotation

Headings and Introductions

The primary aim of a heading is to inform your viewers efficiently about the content they are about to encounter and to orientate themselves within the hierarchy of this content. Main headings typically occupy prominent places in your project's layout, perhaps as a title to introduce a report or at the top of a page or screen.

The suitability of your heading comes down to the language used: what are you going to use this key feature to say? There is no universal practice for what constitutes good use of headings; it will vary considerably between subject areas, project contexts and audience settings. However, I find there are generally four approaches to constructing and using them, as follows.

Statement: These are short headline forms of titles that may highlight a key observation or finding that emerges from a visualisation work. The statement title (Figure 8.1) might be most commonly used with visualisations that offer an explanatory experience, based on the mantra 'if you have something to say, say it'.



Figure 8.1 Examples of 'Statement' Titles

Question: A title presented as a question (Figure 8.2) can offer a compelling way to align your audience's minds with the essence of the curiosity that has driven the project. It prepares them to inherit an appetite to find some notion of an answer to the question posed, which the visualisation should serve. These titles work well for exploratory visualisation experiences.

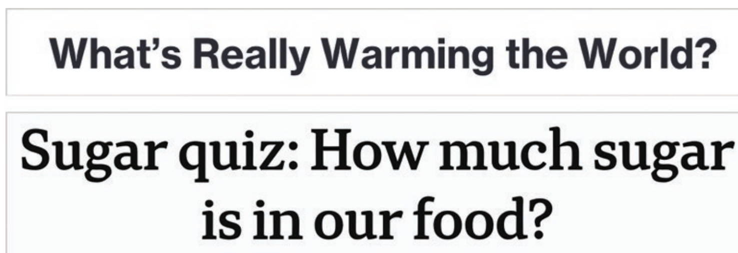


Figure 8.2 Examples of 'Question' Titles

Descriptive: These types of titles generally articulate what is represented on a chart. They are more functional and less editorial in style, but perhaps more informative about what the viewer is about to encounter. Characteristically, descriptive headings (Figure 8.3) tend to be aligned with exhibitory visualisations. This approach would also usually be applied to lower level headings, such as section dividers and localised chart titles, offering details about what each element is about.

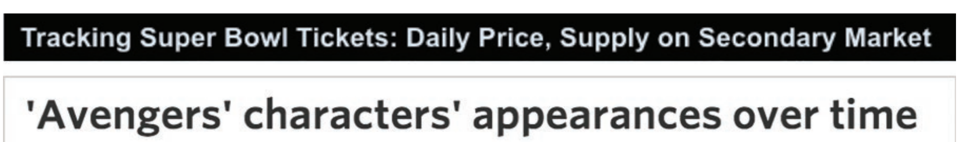


Figure 8.3 Examples of 'Descriptive' Titles

Artistic: This approach tends towards the use of short, succinct and rather enigmatic phrases to convey roughly the nature of the topic, but mainly to pique the curiosity of the audience. The titles are consistent in style with titles typically used for creative endeavours like movies or artworks (Figure 8.4). Due to their punchier size they might be more easily remembered than other approaches. However, they are often necessarily supplemented by more descriptive sub-headings that expands on the detail.

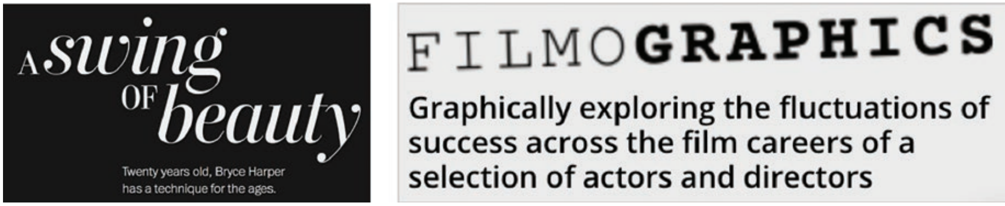


Figure 8.4 Examples of 'Artistic' Titles

Introductions are commonly provided in close proximity to a heading in the form of short paragraphs that concisely explain in further detail what a project is about, why it exists and what it is for. The content of this introduction might usefully explain matters such as:

- details of the reason for the project, perhaps articulating the origin story of the curiosity;
- an explanation of the relevance of this analysis;
- a description of the analysis that is presented;
- a few comments about the main messages or findings the work is about to reveal.

The extracted introduction shown in Figure 8.5 accompanies a main heading to explain the background of the subject, why the analysis has been undertaken, and information about the experts providing their headline observations.

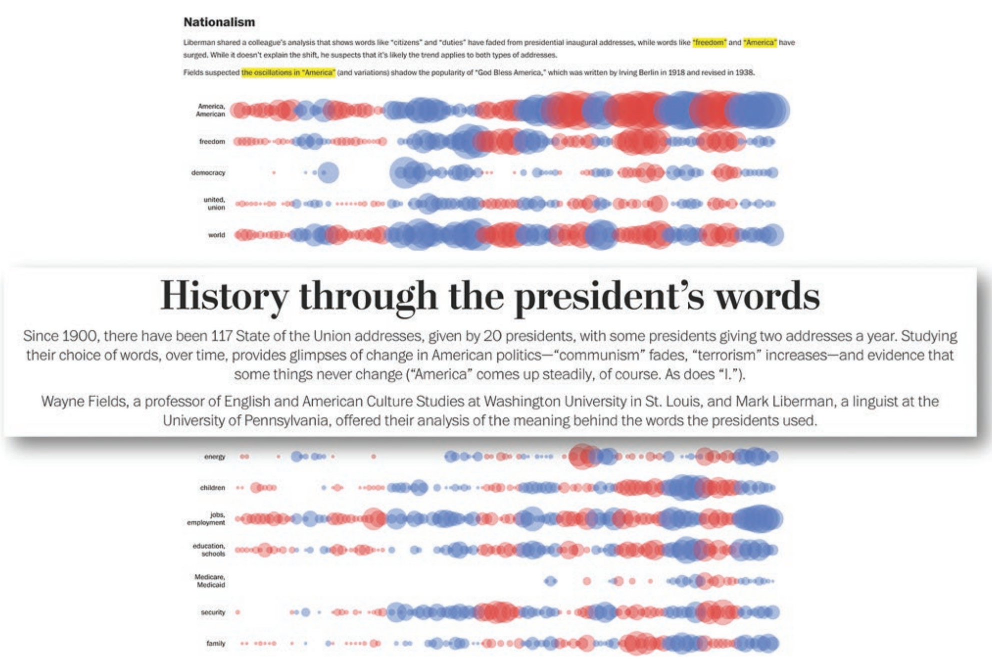


Figure 8.5 Excerpt from History Through the President's Words, by Kennedy Elliott, Ted Mellnik and Richard Johnson (*Washington Post*)

Introductions are, naturally, logically offered near the top or start of a project. Sometimes, through interactivity they are made available on demand through a separate window or pop-up to provide the necessary details. This would be appropriate if they were quite detailed in nature and would otherwise occupy too much precious space. Furthermore, the viewer might encounter the work on a repeated basis, but will only need to read an introduction on the first occasion.

For some projects, the introduction may be used to provide a more extensive description of the data, where data has come from, how it has been transformed, and comments about any assumptions or potential shortcomings.

User Guides

Projects that include features of interactivity may need to offer some level of instruction in the form of prompts or more in-depth user guides. Though the features may not necessarily be overly technical – and easy to learn how to use – instructions can help to enhance the accessibility of your project. In Figure 8.6, there is an input box inviting users to enter a job title. The short instructive sentence explains what to do, and the example text provides clues about the format and phrases you might attempt to search for.

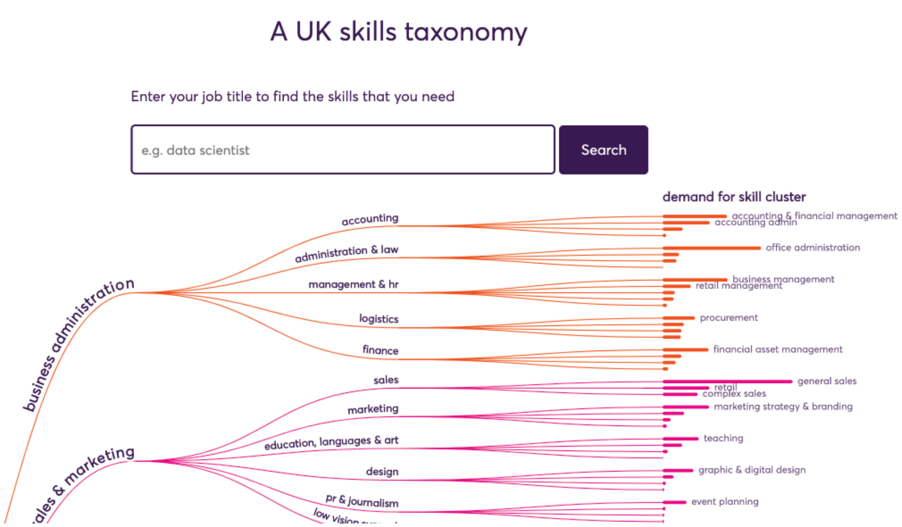


Figure 8.6 Excerpt from Making Sense of Skills: A UK Skills Taxonomy, by Dr Cath Sleeman

It can be a mistake to assume that every user will be sufficiently sophisticated to understand immediately the workings of the functionality you are offering. It can also be a mistake to assume that every user will find all the functions you are offering. A more dedicated user guide that introduces and explains the full repertoire of features might be necessary. Projects like 'Kindred Britain' (Figure 8.7) provide a vast array of means for exploring aspects of

history about the British royal family and aristocracy. Without offering a detailed user guide, many users may miss out on some of the skilfully crafted opportunities to interrogate the data. It is therefore in everyone's interest to provide this type of assistance. Moreover, it is in everyone's interest to provide this assistance using an elegantly presented accessible form, as this example certainly exhibits.

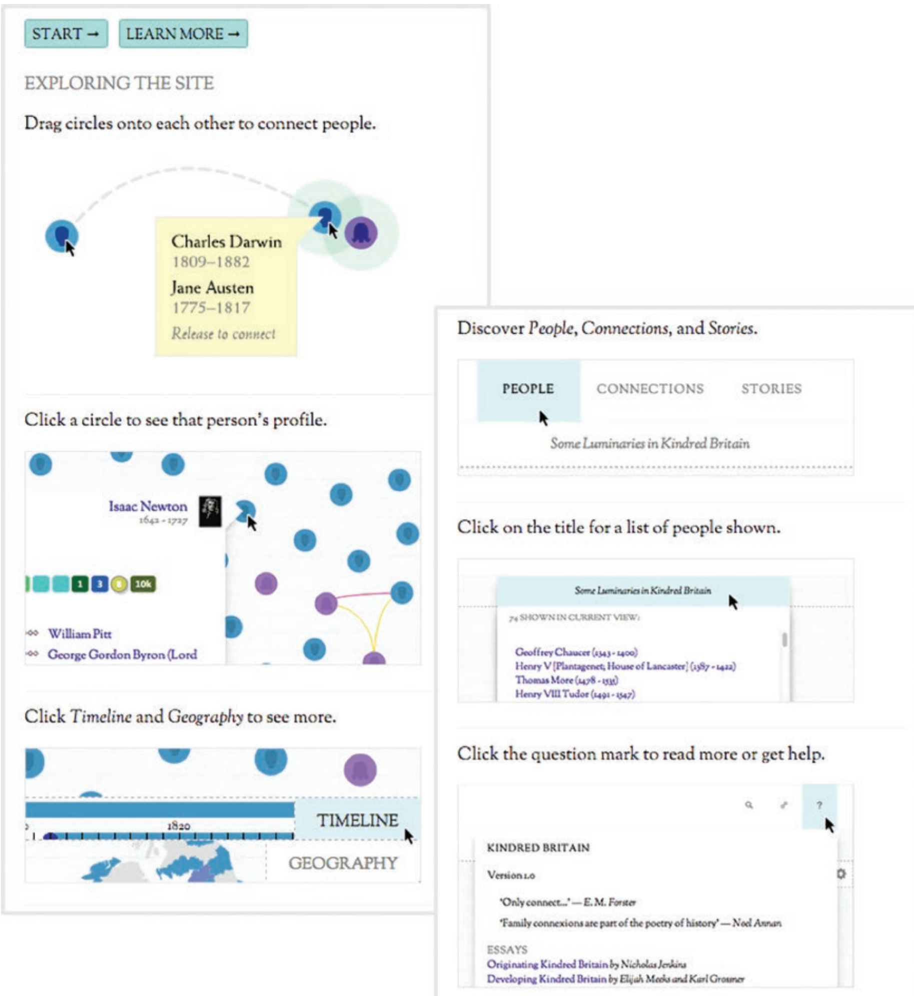


Figure 8.7 Kindred Britain, version 1.0 © 2013 Nicholas Jenkins – designed by Scott Murray, powered by SUL-CIDR

Reader Guides and Legends

Reader guides offer a different type of assistance to user guides in that they focus on helping viewers to understand how to read a chart. If you have used an unfamiliar and/or particularly complicated chart type, with many attributes that need to be decoded and synthesised, you might need to provide instructions to assist viewers who need it.

The first reader guide example, shown in Figure 8.8, comes from work designed by Accurat, a studio renowned for innovative and expressive representation techniques. Given the relative complexity of the encodings used in this piece, it is necessary to equip the viewer with guidance about the layout of each chart panel, what the shading portions and arced lines represent, what the dots and musical notes stand for, and what to imply from the length of the bars.

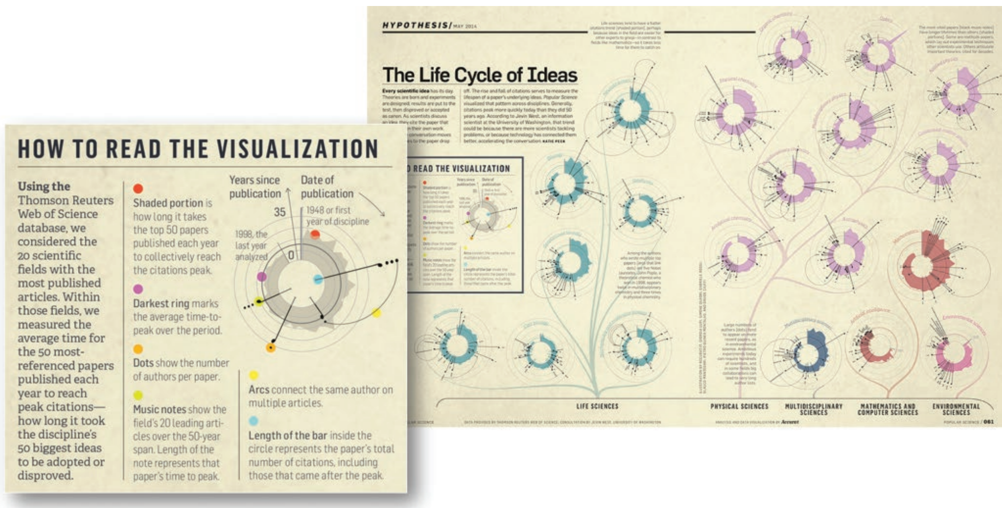


Figure 8.8 The Life Cycle of Ideas, by Accurat

In Figure 8.9, you can see another guide taken from the Gantt chart we saw in the chart gallery, looking at timeline histories of the current and former US national parks. This offers a description about the arrangement of the items, the associations of several elements of symbol and colour usage, as well as definitions of the acronyms used.

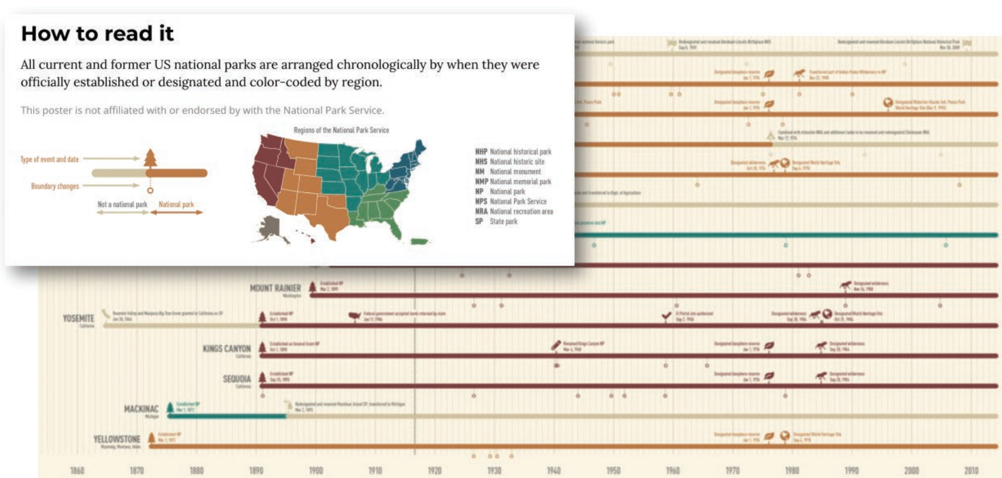


Figure 8.9 Establishment of the US National Parks, by Nicholas Rougeux (www.c82net)

There are many similarities between reader guides and legends. A legend is usually positioned adjacent to a chart and contains one or several keys used to explain associations between categorical attributes or classifications of quantitative scales. Figure 8.10 presents a range of different examples from work you will encounter across this book. The main difference between a legend and a reader guide is that a legend offers far less written explanation, whereas a reader guide more actively coaches a viewer through the reading task.

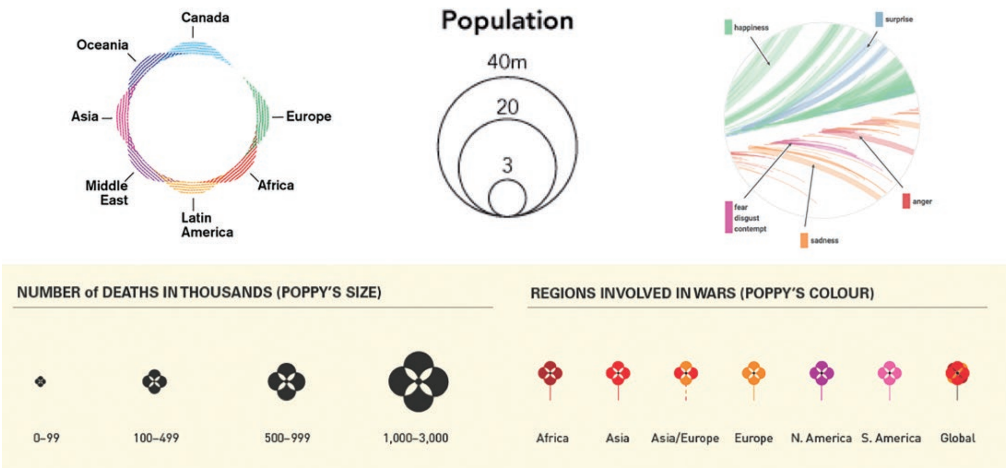


Figure 8.10 Selection of Example Legends

There are different ways of being creative in how you portray a legend. In Figure 8.11 you will see how colour associations are integrated into the introductory text, with certain words highlighted through shading or in the colour of their font. Rather than having a separate colour key elsewhere on the page, this saves space and provides all the setup information and reading instructions in the same place.

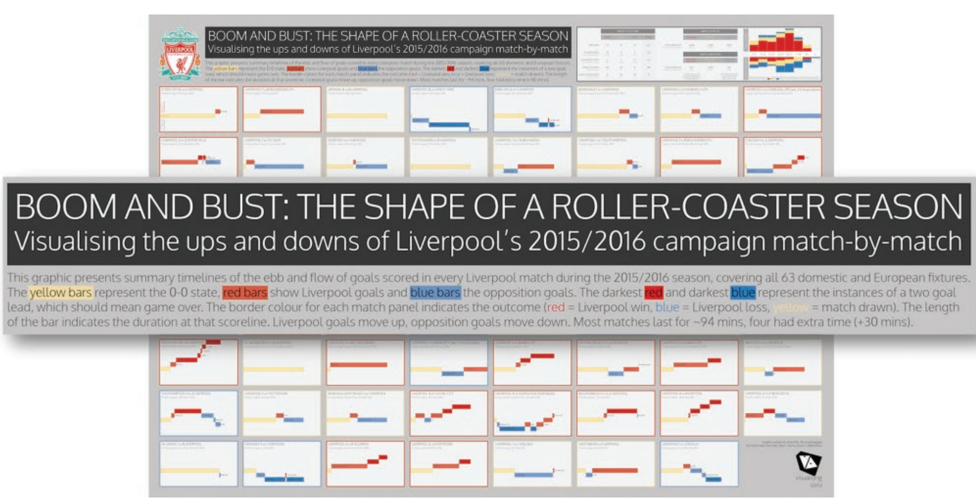


Figure 8.11 Boom and Bust: The Shape of a Roller-coaster Season, by Andy Kirk

In Figure 8.12, you can see a smart way of squeezing more information out of a legend in a visualisation that analyses the language of tweets posted around New York City. The legend takes the form of a bar chart that acts as both a colour key, explaining the association with the different languages, and a method of showing a quantitative summary of the total number of tweets for each language.



Figure 8.12 Twitter NYC: A Multilingual Social City, by James Cheshire, Ed Manley, John Barratt and Oliver O'Brien

Chart Apparatus and References

Chart apparatus relates to the range of structural components found in different chart types, such as axis lines, gridlines or tick marks. I consider these elements of annotation because their role is to help orient viewers in making judgements about size and/or position.

Not all chart types have the same structural apparatus. For example, a pie chart does not have axis lines, a Sankey diagram will not have gridlines. The scatter plot shown in Figure 8.13 is just one example of a work we have seen earlier that includes the full array of chart apparatus. There is no fixed recommended approach on whether to include or exclude these features, but the default treatment applied by most chart applications would generally become quite heavy with the apparatus. Your choices will generally be informed by the degree of emphasis you place on viewers efficiently judging values with precision and also be influenced by your desired presentation style. Mentioning these features in this section is as much about prompting you not just to go through the motions by accepting default thinking.

Beyond the structural elements of chart apparatus, you may find value in incorporating additional markings on your charts to help viewers with the task of interpretation. *Chart references* can be usefully included as visual overlays to provide context of scale, to clarify the expected and unexpected, and to separate the normal from the exceptional. In some ways these features might be considered an extension of data representation, as they will be formed through values of data, but I see them more as annotations focused on providing assistance. There are several different types of references that may be useful to include:

- *Bandings:* These are typically shaded areas that provide some frame of contextual judgement for data values, such as providing a range of historic or expected values. As we saw in the bullet chart in the chart gallery, there are various shaded regions that might help to indicate whether the bar's value should be considered great, good or just average. They might also be used to indicate confidence intervals to convey probabilities or to represent a contextual measurement of margin of error to explain degrees of uncertainty.
- *Markers:* Adding points or symbol markers to a chart might be useful to help indicate statistical features such as comparison against a target, forecast or a previous value of note.
- *Reference lines:* These are useful in chart displays that use position or size along an axis as an attribute for a quantitative value. Line charts or scatter plots are particularly enhanced by the inclusion of reference lines, helping to direct the eye towards calculated trends, constants or averages and, with scatter plots specifically, the lines of best fit or correlation. The example in Figure 8.14 uses a heat map display to show the relative incidence of measles per 100,000 population across each US state over time. The patterns were already indicating a fascinating trend but, by adding the reference line to indicate when the vaccine was introduced in 1963, the compelling story of cause and effect jumps off the page.

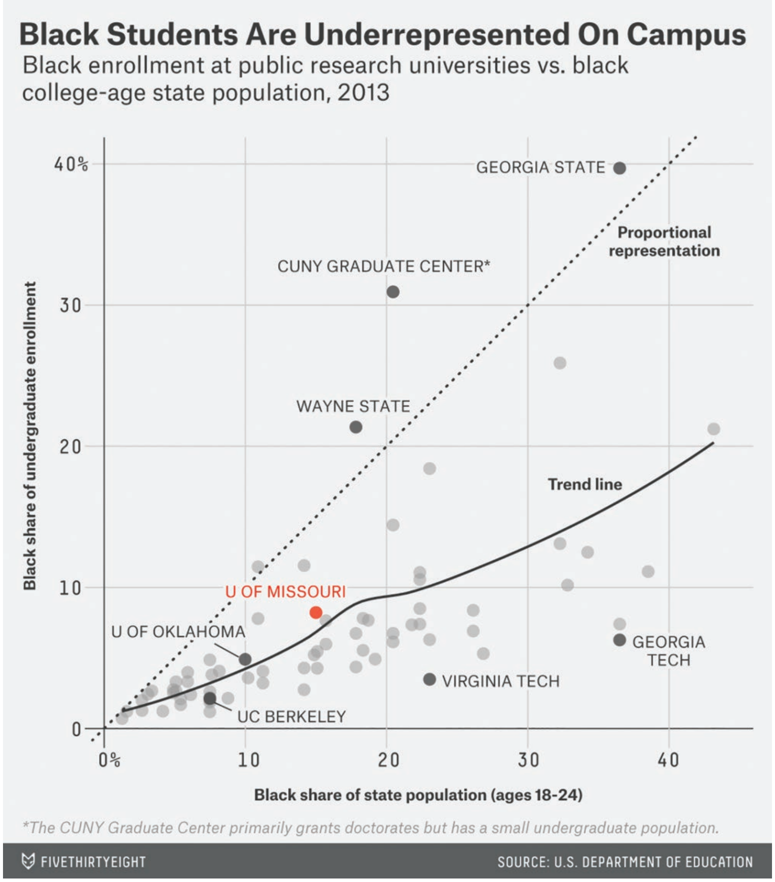


Figure 8.13 Mizzou's Racial Gap Is Typical on College Campuses, by FiveThirtyEight

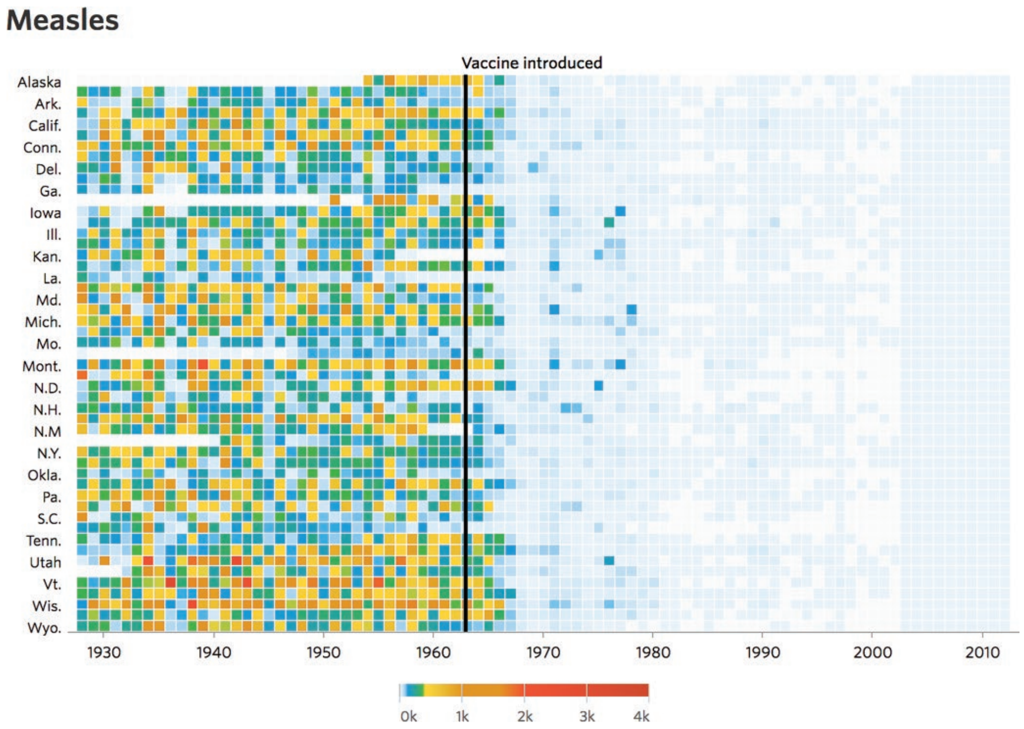


Figure 8.14 Battling Infectious Diseases in the 20th Century: The Impact of Vaccines, by Graphics Department (*Wall Street Journal*)

Chart Labelling and Captions

There are three main features of labelling that you will need to consider adding to your charts, depending on the type of chart you are using. As demonstrated, again, by the chart in Figure 8.13, these include axis titles, axis scales and value labels:

- *Axis titles* describe what values are plotted along each axis. This might be a single word or a short sentence depending on what best fits the needs of your viewers. Often the role of an axis is already explained (or implied) by headings or introductions, but do not always assume this will be automatically clear to your viewers.
- *Axis scales* provide references along each axis to identify the categorical items, quantitative value intervals or the dates with a time frame. For categorical axes (as seen in bar charts and heat maps) one of the main judgements relates to the readability of the label and how well it fits into the space you have. For non-categorical data the main judgement will be what quantitative intervals to use. This will be shaped by considering the most relevant interval for the subject matter and the required precision in readability. It can also come down to what offers the most elegant rhythm in your display – does it feel too fussy or too sparse? Depending on your editorial framing definitions, you might

also expand your quantitative range outside the observed value range in order to use empty chart space to support a point of narrative.

- *Value labels* will appear in proximity to specific mark encodings inside a chart. Typically, these labels will be used to reveal a quantity, such as showing the percentage sizes of the sectors in a pie chart or the size of categorical bars. Having the option to reveal values through interactive annotations can be a nice option, as it reduces clutter from a display. Unless you are highlighting important values for key items, if you have clear axis labels you should not need to double up your value reading assistance. Choose one or the other.

Captions are often included in explanatory projects to offer more localised comments that transcend the general role of an introduction, user or reader guide. The excerpt shown in Figure 8.15 comes from an interactive project that offers users scrollable navigation through discrete sequences of analysis about the history and growth of the #MeToo movement. At each milestone stage, captions are displayed to offer historical narrative and context to where we are in the story. Furthermore, more detailed value labels emerge, in the form of significant tweets, showcasing some of the most influential actions, actors and conversations in this campaign.



Figure 8.15 MeTooMentum, by Valentina D'Efilippo (design) and Lucia Kocincova (development)

In 'US Gun Deaths' (Figure 8.16), there is a clever feature that combines annotated captions with interactive data adjustments. Below the main chart there is a 'What This Data Reveals' section which presents some of the main findings. The captions double up as clickable shortcuts that, when selected, quickly apply the relevant filters to the main display so users can see in the chart the relevant data that supports each commentary.



Figure 8.16 US Gun Deaths, by Periscopic

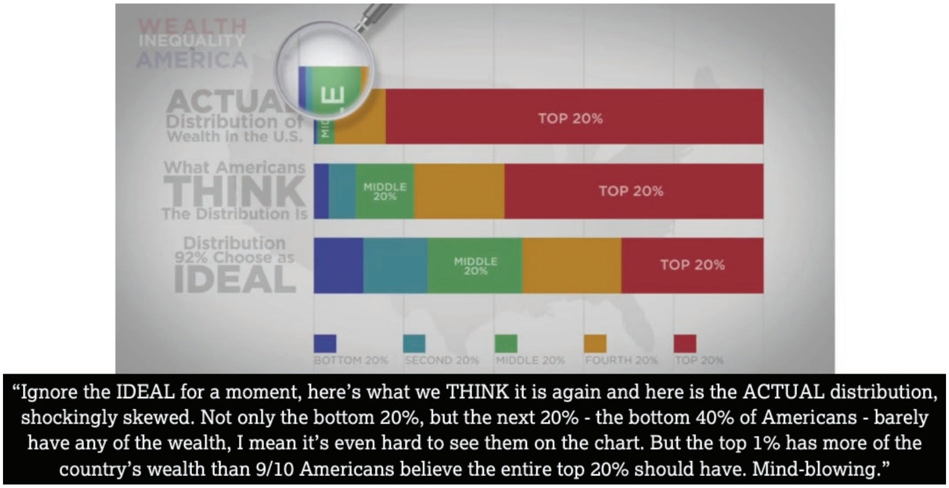


Figure 8.17 Excerpt from Wealth Inequality in America, by YouTube user 'Politizane'

As creative tools become more ubiquitous, there are new opportunities for incorporating audio as a means for verbally narrating a subject and explaining key messages. One of the standout projects using this approach in recent years was the video 'Wealth Inequality in America' (Figure 8.17), as mentioned in Chapter 3. In this videographic, the voiceover provides a compelling and cohesive narrative about the subject with the visuals supporting what is being described.

Footnotes and Methods

Often the final visible feature of your display, *footnotes* provide a convenient place to share useful information that further substantiates the transparency of your work:

- *Data sources* from where your raw material was acquired should always be provided, ideally in close proximity to the relevant charts if several are to be included.
- *Credits* will list the authors and main contributors of the work, often including details about methods of contact for further information or feedback.
- *Attribution* is important if you wish to recognise the influence of other people's work in shaping your ideas or, for instance, to acknowledge your use of open source applications or typeface.
- *Usage* information might explain the circumstances in which the work can be viewed or reused, whether there are any confidentialities or copyrights involved.
- *Time/date stamps* are useful to include so they indicate the moment of production/publication and from which it will be possible to determine the work's ongoing accuracy and relevance.

Annotation is one of the most important aids to ensure you secure and sustain trust from your viewers by demonstrating integrity and openness. If you have to undertake a lot of work to transform your data for use in your analysis, it may be necessary to provide more detail using a *methods* statement, an example of which is shown in Figure 8.18. In the spirit of 'show your workings', these sections will typically extend beyond comments about the sources and methods of collecting data to explain any assumptions that have been made, details of calculations applied, the criteria for editorially framing the work (inclusions/ exclusions), and maybe what imperfections existed in the data and how you have handled them.

DATA SOURCES AND TREATMENT

Up to and including 2008 the data was sourced from the databaseOlympics website. Subsequent data was sourced from the BBC Olympics websites.

The criteria for the inclusion of events was based on selecting:

- Only those events where an absolute time-based measurement determines the medal awards.
- Only those events for which there is a final race, over a set distance (eliminating events such as the Cycling pursuit).
- Only events that were part of the 2016 Rio Olympics.
- Only events that have been part of at least the previous 3 summer Olympics.

This reduces the qualifying sports to athletics (track), canoeing, rowing and swimming and eliminates some of the events within each. Some data cleaning has been conducted on missing values. Note that there are some very early events for which no more reliable times can be sourced.

Figure 8.18 The Pursuit of Faster, by Andy Kirk and Andrew Witherley

8.2 Influencing Factors and Considerations

You have now been introduced to the roles of different annotation options, so how do you decide what types of annotations to incorporate and what level of assistance to offer your audience? Let's look at some of the main factors and considerations that will influence your decisions.

Audience: Given that most annotations serve the purpose of viewer assistance, your decisions will inevitably be influenced by the characteristics of your intended audience. Having an appreciation of and empathy towards the knowledge and capabilities of the different cohorts of viewers is of principal concern. It can be hard to find a solution that suits all, especially if your viewers are diverse, but here are some of the main issues to consider:

- *Subject:* How well acquainted are they with your project's subject matter? Will they understand what the data is about? Will they understand language, such as specific terminology, acronyms or abbreviations?
- *Perceiving:* How familiar are they with the chart type(s) you have used? If they are unfamiliar, is it easily learnable or will they need some explanations?
- *Interpreting:* Will they know how to interpret the meaning of what is presented? Do they need help in determining what features are good or bad, significant or insignificant?
- *Interactive functions:* How confident will they be in understanding how to use different features of interactivity?

Setting: Providing a sufficient amount of assistance is about balance: too much assistance makes the annotations included feel overburdening; too little and there is too much scope for misconceptions and misinterpretation to prosper.

> 'Think of the reader – a specific reader, like a friend who's curious but a novice to the subject and to data-viz – when designing the graphic. That helps. And I rely pretty heavily on that introductory text that runs with each graphic – about 100 words, usually, that should give the new-to-the-subject reader enough background to understand why this graphic is worth engaging with and sets them up to understand and contextualize the takeaway. And annotate the graphic itself. If there's a particular point you want the reader to understand, make it! Explicitly!' **Katie Peek, Visualisation Designer and Science Journalist, on making complex and/or complicated subject matter accessible and interesting to her audience**

A setting that is consistent with the need to deliver immediate insights will need annotations to help fulfil this rapid exchange of understanding. There will be no time for long introductions or patience with explanations about how to read charts. Conversely, a visualisation about subject matter that is inherently complex may warrant more assistance and invite the viewer to embark on a process of learning. If there is time for a viewer to engage with a user or reader guide, it is entirely reasonable to include such features since the rewards should outweigh the efforts involved.

Purpose: Your definitions for the intended *tone* and *experience* of your work will influence the type and extent of annotation

features required. If you are working on a solution that leans more towards the 'reading' tone, you are placing an emphasis on the perceptibility of the data values. It therefore makes sense that you should aim to provide as much assistance as possible (especially through extensive chart annotations) to maximise the efficiency and precision of this reading process. In combination with your editorial 'focus' thinking, you might choose to emphasise specific items over others, in which case: What are the criteria and reasoning for this? Will the selective labelling be fixed, or should it be driven by interactive selection?

If you are providing an 'explanatory' experience it would be logical to employ as many devices as possible that will help inform your viewers about how to read the charts (assisting with the 'perceiving' phase of understanding) and also bring some of the key insights to the surface, making clear the meaning of the quantities and relationships displayed (thus assisting with the 'interpreting' phase). The use of captions and reference markings will be particularly helpful in enabling this.

'Exploratory' experiences are likely to need instructive guides, ensuring that viewers (or specifically, in this case, users) have as much understanding as possible about the functional controls available. Features like reader guides, chart apparatus and labelling will still be relevant, irrespective of the intended experience. Characteristically, 'exhibitory' work includes only chart-level annotation, as it is more about providing a visual display of the data rather than offering explanatory insights or instructions for exploratory interrogation. The assumptions with exhibitory experiences are that the audience do not require extensive assistance.

Accessible design: Although equally connected with concerns about elegance, decisions on typography will have an influence on the accessibility of your work. As you will have observed, many features of annotation utilise text. This means your decisions will be concerned not just with *what* text to include, but also with *how* it will look. Typography is another element of data visualisation design thinking that exists as a significant subject in its own right, but here is a short guide to inform your thinking:

- A *typeface* is the styled glyphs representing individual letters, numbers and other symbols. Tahoma and Century Gothic are different typefaces. A typeface can have one or many different fonts in its family.
- *Fonts* are variations in the dimension of your typeface, such as weight, size, condensation and italicisation. This *font* and this **font** both belong to the **Georgia** typeface family but display variations in size, weight and italicisation.
- Serif typefaces are characterised by extra little flourishes in the form of a small line at the end of the stroke in a letter or symbol. Garamond is an example of a serif font. Serif typefaces are generally considered to be easier to read for long sequences of text (such as the full body text) and are especially used in print displays.
- Sans-serif typefaces have no extra line extending the stroke for each character. Verdana is an example of a sans-serif typeface. These typefaces are commonly used for shorter sections of text, such as axis or value labels or titles, and for screen displays.

Your typeface and font choices should be based on optimising the legibility and meaning of text elements across your display:

- In terms of legibility, viewers need to be able to read the words and numbers on display without difficulty. Quite obvious, really. Some typefaces (and specifically fonts) are more easily read than others. Some are better applied to help make numbers clearly readable, others work better for words and passages or sentences.
- Just as variation in colour implies meaning, so does variation in typeface and font. If you make some text capitalised, large and bold-weight, this will suggest it carries greater significance and portrays a higher prominence across the object hierarchy than text presented in lower case, with a smaller size and thinner weight. In general, you should seek to limit the variation in font where possible and only vary it when the property it is applied to needs to be distinguishable from other properties.

Deciding on the most suitable choice of typeface and variety of font is something that will ultimately come down to experience and being influenced by other creative work you encounter. We all have our own preferences but, in practice, I find most typographic decisions I make rely on experimentation.

Elegant design: A final judgement about annotations concerns avoiding the potential clutter and obstruction caused by them. Any annotation feature included in your work will add more content. These features have to be located somewhere. Too much and the display becomes cluttered, overwhelming and potentially undermines the intention of being helpful; too little and viewers may be faced with the demands of working things out themselves, when assistance would make that prospect far easier.

Summary: Annotation

Features of Annotation

This chapter described the importance of providing useful assistance to your viewers, introducing some of the many helpful features to consider, including:

- Headings and introductions: Titles, subtitles and section headings, often combined with longer passages to describe the background and aims of a project.
- User guides: Advice or instructions on how to use interactive features.
- Reader guides and legends: Detailed instructions advising viewers how to perceive and interpret the chart, describing the associations between data values and attribute classifications.
- Chart apparatus and references: Structural components found in different chart types, such as axis lines, gridlines or tick marks, as well as markings that assist with interpretation.
- Chart labelling and captions: Axis titles, axis labels, value labels and commentaries.
- Footnotes and methods: Include data sources, credits, and time/date stamps. May be expanded to provide more detailed description of data handling processes, assumptions and shortcomings.

Influencing Factors and Considerations

If these were the options, how did you make your choices? The influencing factors included:

- Audience: Considering the characteristics and needs of the audience to determine what assistance they might need.
- Setting: Will the audience have the scope to engage with annotations if the encounter is characterised by time pressures?
- Purpose: The tone and experience offered will influence the type of annotations required.
- Accessible design: Many annotations are based on text displays and so you need to consider the legibility of the typeface you choose and the logic behind the font–size hierarchy you display.
- Elegant design: Minimise the clutter.

General Tips and Tactics

- Attention to detail is imperative. All introductory information, project instructions, captions and value labels need to be accurate. Always spell-check digitally and manually and ask others to proofread if you are too 'close' to the work to see it rationally.
- Testing with sample members of an audience may save you a lot of pain by intercepting any shortcomings or excesses in the annotations you plan to offer.

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

9
Colour

Having now established the charts you intend to use, the interactive features that might be required and the elements of annotation that will be useful, you have effectively selected all the visible contents of your visualisation. The remaining two layers of design thinking are concerned with the appearance of these contents. In the final chapter after this we will consider decisions about composition, but before that we look at the critical issue of making choices about colour.

Colour is a most potent visual stimulus. The choices we make will have an immediate impact on the eye of the viewer, offering sensory cues about the meaning and organisation of a display. Variation in colour implies significance. When a colour looks like it conveys meaning, the viewer will think about that and spend time establishing what the meaning is. Many of the chart types employ an attribute of colour to represent data values. Whether it is used to classify quantitative scales or to associate with discrete categorical values, there is a lot riding on your colour choices being astutely judged.

The chapter opens with an overview of colour models, offering a foundation for your understanding about this topic. After that you will learn about the different ways and places in which colour is to be used, starting from inside a chart and then working outwards across the rest of the visualisation anatomy. Your objective is to establish meaning first and worry about decoration last. As before, you will then reflect on the main factors that will ultimately shape your choices.

9.1 Overview of Colour Models

Colour is a vast theoretical subject rooted in the science of optics – the branch of physics concerned with the behaviour and properties of light – as well as colorimetry – the science and technology used to quantify and describe human colour perception. The challenge of writing about colour in the context of visualisation thinking is to establish a pragmatic cut-off point that avoids sinking too deeply into these sciences, but still provides a rigorous basis for the recommendations that follow.

The most relevant starting point for this overview is to recognise that, when dealing with issues of colour in data visualisation, you will almost always be creating work using a computer. There

are exceptions of course, as we have seen in work created using Play-Doh and colouring pencils, but mostly you will be using software viewed through an electronic display.

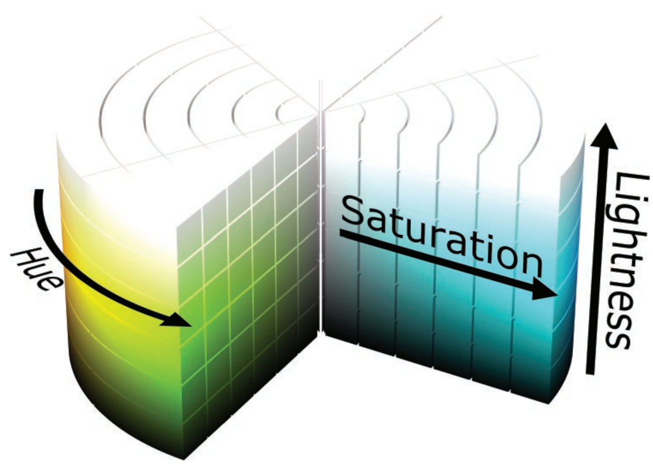
A discussion about colour theory needs to be framed around the RGB (Red, Blue, Green) colour model. This is used to define the combination of light that forms the colours you see on a screen, conceptually laid out in a cubic space based on variations across these three attributes. Even if you create work with tools that use hexadecimal codes to specify your colour choices, these specifications are still based on a mix of red, green and blue light. The 'hex' values take the form of #RRGGBB using two-digit codes for each component ranging from 00 to FF.

The output format of your work will vary between screen display and print display. If you are creating something for print you will shift your colour output settings to CMYK (Cyan, Magenta, Yellow and Black). This is the model used to define the proportions of inks that make up a printed colour. This is known as a subtractive model, which means that combining all four inks produces black. RGB is an additive model as the three screen colours combine to produce white.

CMYK communicates from your software to a printer, telling it what colours to print as an output. RGB does the same but communicates the colour messages to a screen display. Neither of these, though, really offer a logical model for us to think about the choices we might make on which colours to use in a visualisation design. We require an alternative model of colour thinking.

One of the most accessible colour models for considering the application of colour in data visualisation is known as HSL (Hue, Saturation, Lightness), and was devised by Albert Munsell at the start of the 20th century. These three dimensions (Figure 9.1) combine to make up what is known as a cylindrical-coordinate colour representation of the RGB colour model.

Figure 9.1 HSL Colour Cylinder: Image from Wikimedia Commons published under the Creative Commons Attribution-Share Alike 3.0 Unported license



Hue is considered the *true* colour. When you are describing or labelling colours you are most commonly referring to their hue: think of colours of the rainbow ranging through mixtures of red, orange, yellow, green, blue, indigo and violet (Figure 9.2). Hue is considered a qualitative colour attribute because it is defined by difference and not by scale. With hue there are no shades (adding black), tints (adding whites) or tones (adding grey).



Figure 9.2 The Colour 'Hue' Spectrum

Saturation defines the purity or colourfulness of a hue. This does convey a scale ranging from intense, pure colour (high saturation) through increasing tones to what is technically the 'no-colour' state of grey (low saturation) (Figure 9.3). In language terms think *vivid* through to *muted*.

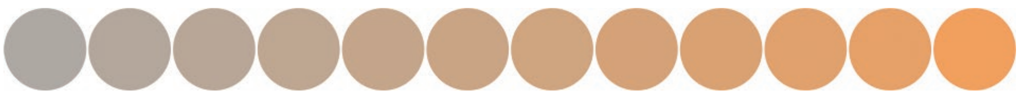


Figure 9.3 The Colour 'Saturation' Spectrum

Lightness defines the contrast of a single hue from dark to light (Figure 9.4). It is not a measure of brightness – there are other models that define that – rather a scale of light tints (adding white) through to dark shades (adding black). In language terms I actually tend to think of lightness more in terms of degrees of darkness, but this is just a personal mindset. Also note that, technically speaking, black, white and grey are not considered colours, but for the sake of this chapter we will continue to think of them as being so.

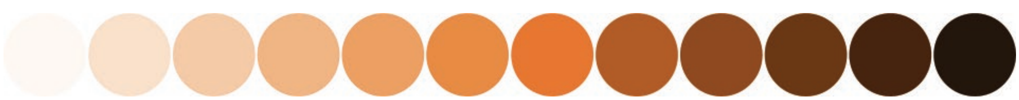


Figure 9.4 The Colour 'Lightness' Spectrum

Alternative models exist offering variations on a similar theme, such as HSV (Hue, Saturation, Value), HSI (Hue, Saturation, Intensity), HSB (Hue, Saturation, Brightness) and HCL (Hue, Chroma, Luminance). These are representations of the RGB model space but involve different mathematical translations into and from RGB. They also bring differences in the meaning of the same terms (definitions of hue and saturation vary local to each model). The biggest difference, though, concerns whether the models specify colour from the perspective of its *quality* (as in how a colour is intended to appear) or as it is *perceived* (as in how a colour is ultimately experienced). Pantone is another colour space that you might recognise. It offers a proprietary colour-matching, identifying and communicating service for print, essentially giving 'names' to colours based on the CMYK process.

There are arguments against defining colour thinking using the HSL model. While it is fine for colour setting (i.e. an intuitive way to think about and specify the colours you want to set in your visualisation work), the resulting colours will not be uniformly perceived the same, from

one device to the next. This is because there are many variables that affect the projection of light when displaying colour, which means the same perceptual experience will not be guaranteed.

Some make a case for other models, such as CIELAB, as being more rigorous in the way they offer an absolute, rather than relative, definition of colour for both input and output. Though I understand the rationale, models like this can become too detached from the ideal pragmatism of conceiving appropriate colour choices for visualisation design thinking. For the purpose of this chapter, I will therefore draw suggestions from the HSL model.

9.2 Features of Colour

Data Legibility

Data legibility concerns the astute use of colour attributes to represent data values. The term *legibility* places an emphasis on making sure the differences between and associations of any colours used are readable and meaningful. There are different ways of optimally using colour to represent values, depending on whether they are showing quantitative or categorical data.

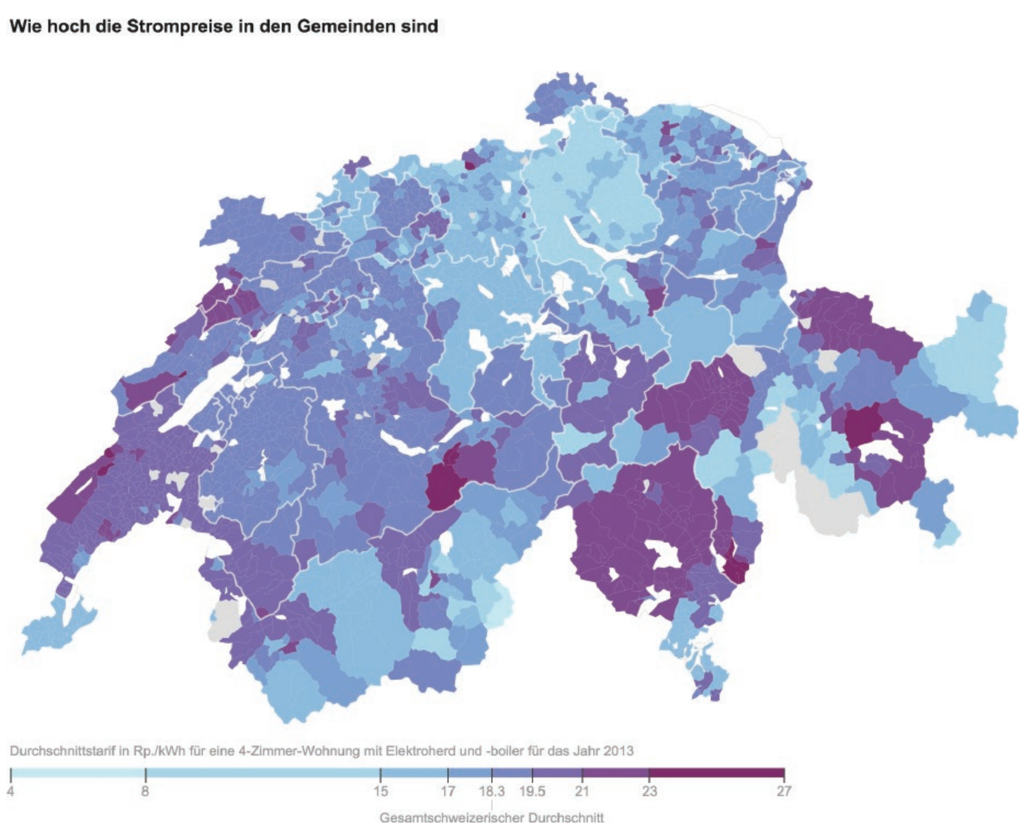


Figure 9.5 What are the Current Electricity Prices in Switzerland? [Translated], by Interactive things for NZZ

Colouring quantitative scales: When using colour to classify quantitative values the primary aim is to create a sufficiently intuitive scale that facilitates an understanding of the hierarchy of values. Variation in the *lightness* of a hue is typically the approach used for differentiating quantities.

The viewer should be able to discern, at least, whether a particular colour represents a larger or smaller quantitative value than another quantity. Assessing the relative contrast between two colours is generally how we construct a quantitative hierarchy. Absolute judgements can be harder, even with a colour key provided for reference, and especially if you employ a continuous gradient scale. The visual system is not always capable of reliably matching exact differences in colour.

To maximise the efficiency of judging absolute values, quantitative colour scales are often divided up into discrete classes, with each increasing in shade (towards dark) or tints (towards white). This helps viewers detect local variations of colour. In the choropleth map in Figure 9.5, showing the variation in electricity prices across Switzerland, the darker shades of blue indicate the higher values, the lighter tints the lower prices.

Similarly, there are fascinating patterns that emerge in the next map (Figure 9.6), comparing increases in the percentage of people gaining health insurance in the USA during 2013–14. The data is broken down to county-level detail with a colour scale showing a darker red for the higher percentage increases.

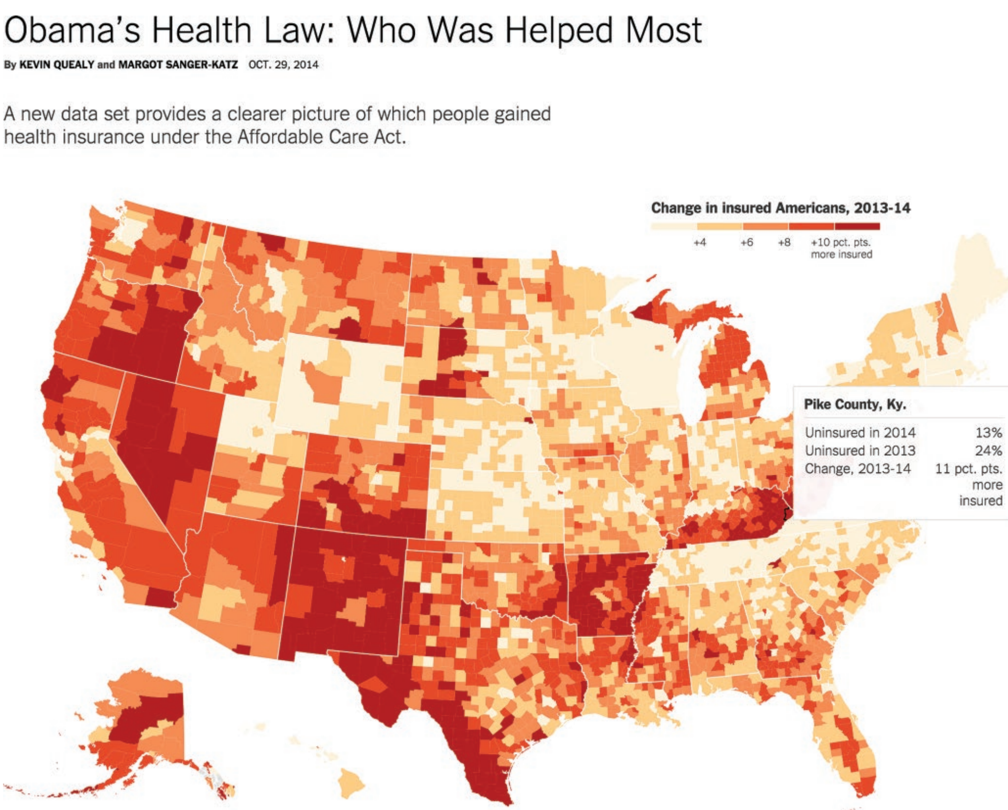


Figure 9.6 Obama's Health Law: Who Was Helped Most, by Kevin Quealy and Margot Sanger-Katz (*New York Times*)

Aside from the big-picture observations of the darker shades in the west and the noticeably lighter tints to the east and parts of the mid-west, take a closer look at some of the interesting differences at a more local level. Notice the stark contrast across state lines between the dark regions of southern Kentucky (to the left of the annotated caption) and the light regions in the neighbouring counties of northern Tennessee. Despite their spatial proximity, there are clearly strong differences in enrolment on the programme among residents of these regions.

These two examples both employ a *converging* colour scale, moving through discrete variations in the lightness of a single hue to represent small through to large quantities. Sometimes the shape and range of your data may warrant a *diverging* colour scale. This is when you want to show how quantities are changing in two directions either side of a specified breakpoint. This breakpoint is commonly set to separate values visually above or below zero or those either side of a meaningful threshold, such as a target, an average or a middle value.

The map in Figure 9.7 demonstrates this approach, through plotting data about the demographics of the neighbourhoods around Berlin with a specific focus on the proportions of inhabitants who are new or native Berliners. There is a diverging colour scheme used to indicate whether there is a dominance of new Berliners (light orange to dark) or native Berliners (light blue to dark). As this work is also interactive, the readability of the colour scales is supplemented by annotated tooltips presenting the actual values in each location.

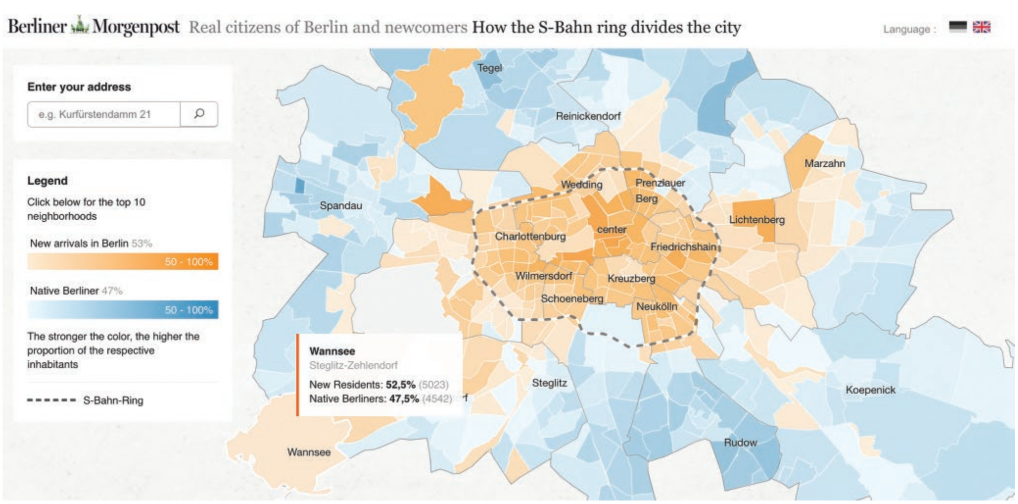


Figure 9.7 Native and New Berliners – How the S-Bahn Ring Divides the City, by Julius Tröger, André Pätzold, David Wendler (*Berliner Morgenpost*) and Moritz Klack (webkid.io)

Although entirely continuous colour scales are not uncommon, usually there is value in dividing up your converging or diverging scales into discrete classes. This needs careful thought to ensure you get the right balance between aiding judgements of the relative order of magnitude as well as the absolute magnitudes. There are two key factors to consider when judging your scales:

- Are you plotting *observed* data or *observable* data? You might have data based on responses to a survey that measures levels of satisfaction. The values in your dataset range from 42% to 82%. This is the observed data. However, it was possible for these responses to have ranged from 0% to 100%, so will your colour classifications be based on the observed range or on the potentially observable data range?
- What is the distribution of your data? Does it make sense to create equal intervals for your colour classifications or are there more meaningful divisions that better reflect the shape of your data? Sometimes, you will have legitimate outliers that, if included, will stretch your colour scales far beyond the meaningful concentration of where most values reside.
- For diverging scales, the respective colour classification increments either side of a breakpoint need to imply the same quantitative increment in both directions. For example, if you use a shade of colour to represent +10% on one side of the breakpoint, the respective colour shade for –10% on the other side of the breakpoint should have the same shade intensity but for a different hue.
- Additionally, the darkest shades of hues at the extreme ends of a diverging scale must still be discernible. Sometimes darkest shades will be too close to black and viewers will no longer be able to distinguish differences in the underlying hue.

One of the common mistakes in using colour to represent quantitative data is in the use of the much-derided rainbow scale. The map in Figure 9.8 shows some particular alarming high temperatures across Australia. Consider the colour key to the right of the map. Is this a sufficiently intuitive scale to identify quantitative classifications? If there were no key provided, would you still be able to perceive the order of magnitude relationship between the colours

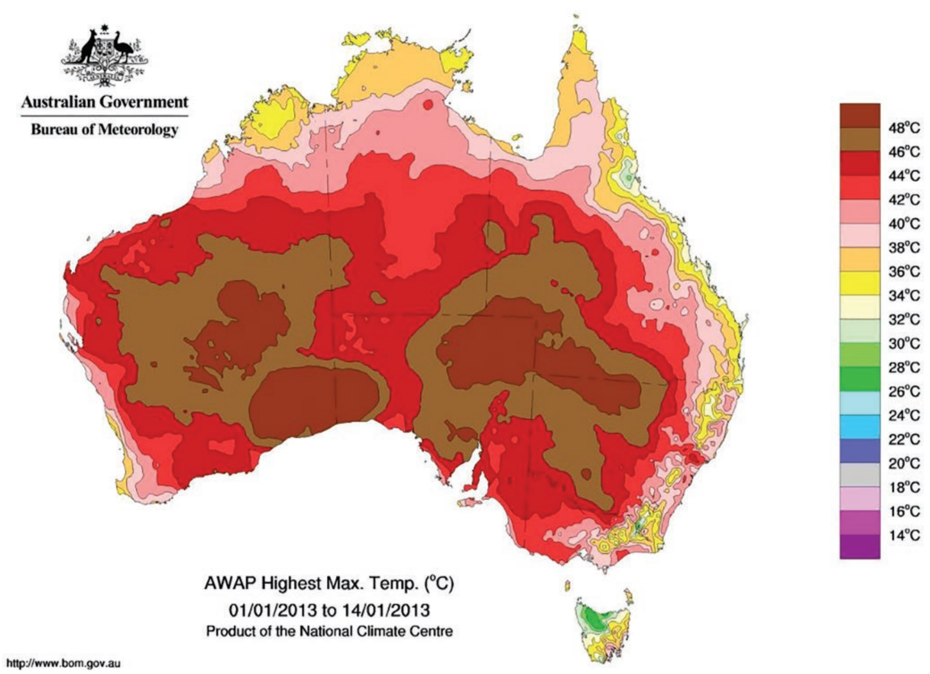


Figure 9.8 Highest Max Temperatures in Australia, produced by the Australian Government Bureau of Meteorology

on the map? If you saw a purple colour next to a blue colour, which would you expect to mean hotter and which colder?

While the general implication of blue = 'colder' to red = 'hotter' is represented in parts of this temperature scale, the presence of other hues obstructs the accessibility and creates inconsistency in logic. For instance, do the colours used to show 24°C (light blue) jumping to 26°C (dark green) make sense? How about 18°C (grey) to 20°C (dark blue), or the choice of the mid-brown used for 46°C which interrupts the sequence of red shades? If you saw on the map a region with the pink tone, as used for 16°C, would you be confident that you could easily distinguish this from the lighter pink used to represent 38°C? Unless there are meaningful thresholds within your quantitative data – justifiable breakpoints – you should only vary your colour scales through the lightness dimension, not the hue dimension.

Colouring categorical classifications: When using colour to classify nominal categories the primary aim is to create a clear, visible distinction between each unique categorical association. The viewer should be able to discern different values as efficiently and accurately as possible. You are *not* seeking to imply any notion of order or magnitude. You just want to help differentiate each category from the others in a way that preserves a sense of equity among the colours deployed.

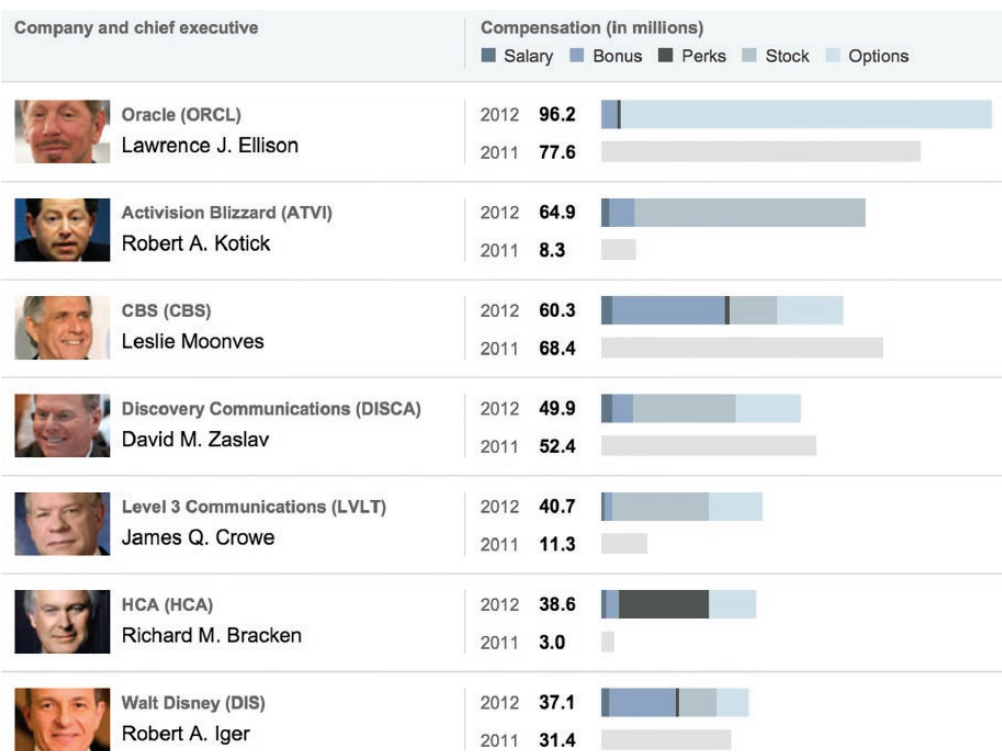


Figure 9.9 Executive Pay by the Numbers, by Karl Russell (*New York Times*)

Variation in hue is typically the colour dimension to consider using for differentiating categories. From a stylistic perspective, you might choose to vary the saturation across all hues, but you should not consider using variations in the lightness dimension. As you can see demonstrated in Figure 9.9, the lightness variation of a blue tone makes it quite hard to connect the colour associations presented in the key at the top with the colours displayed in the stacked bars underneath. With the shading in the column header and the 2011 grey bar also contributing similar tones, the viewer's visual processing system has to work much harder to determine the associations than it should need to do.

Often the quantity of distinct categories you will need to differentiate between using colour will be relatively few in number. In Figure 9.10, two colours are used to separate and associate the panels of analysis in the charts showing margins of victories across all Olympic 100m events for women and men respectively.

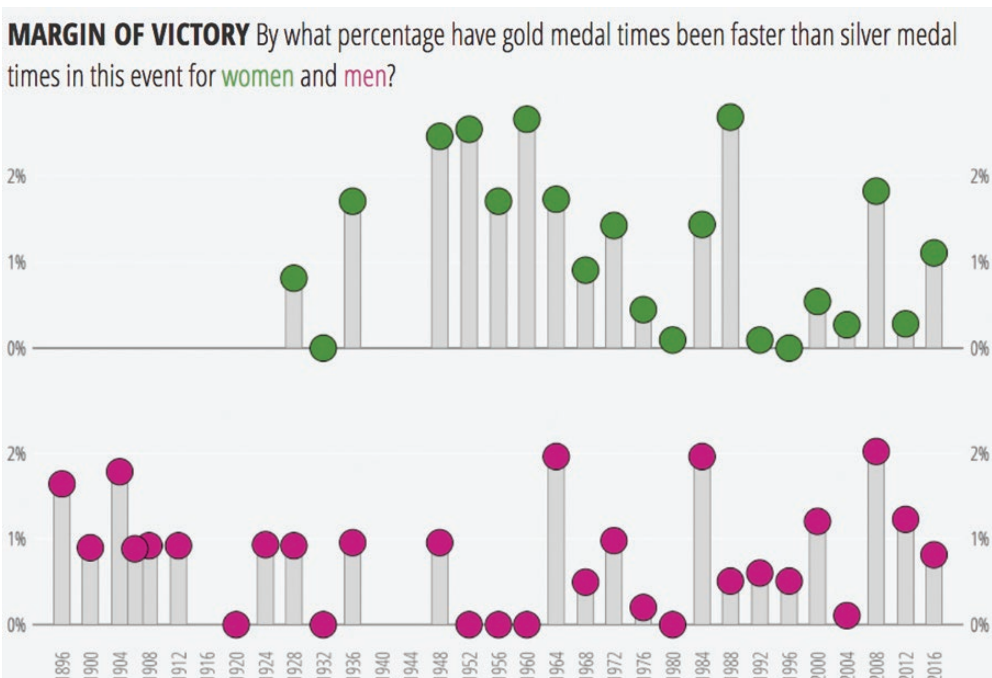
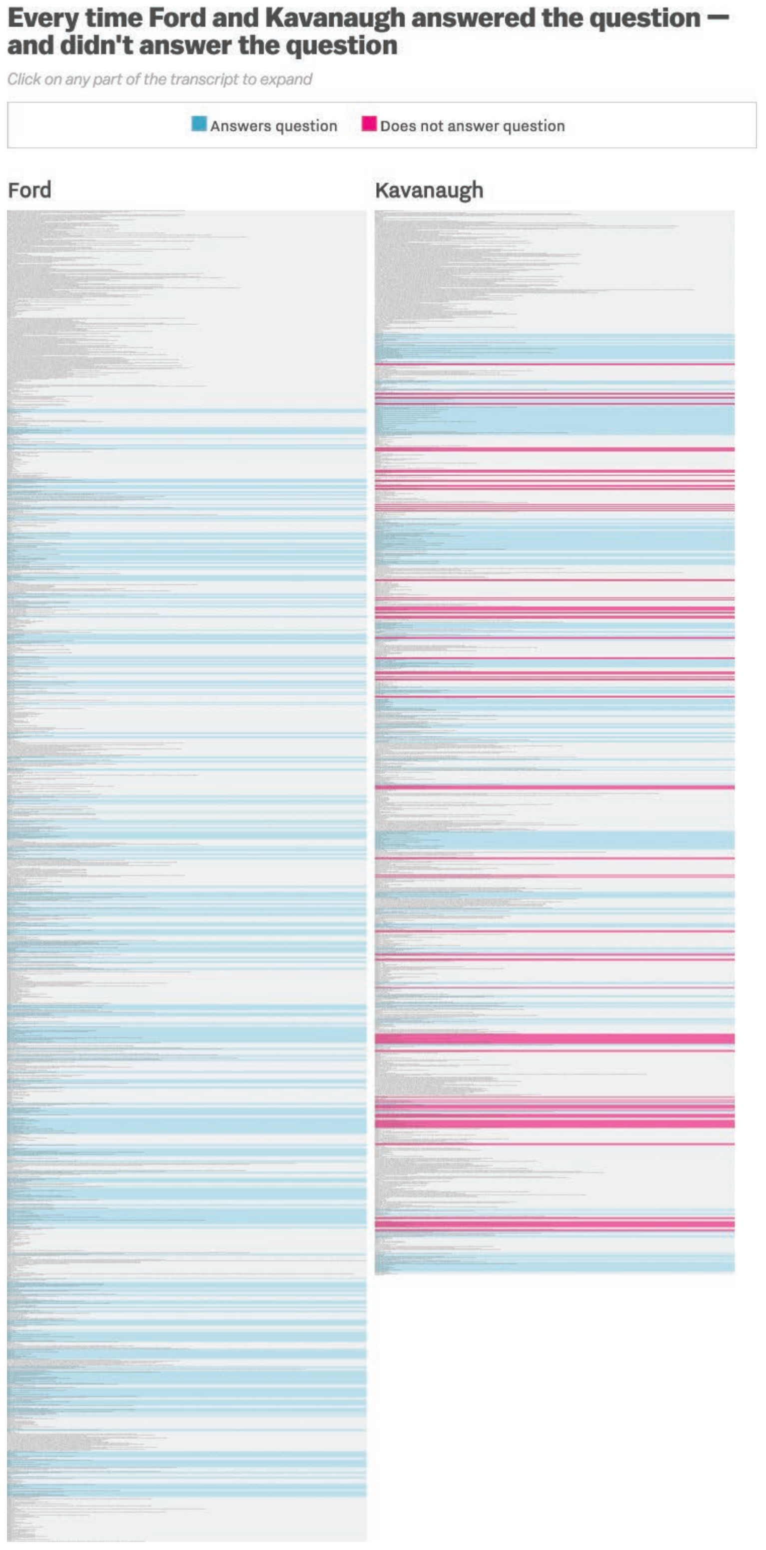


Figure 9.10 The Pursuit of Faster, by Andy Kirk and Andrew Witherley

In Figure 9.11, colours are used to represent key moments from the transcript of the Senate testimony of then Supreme Court nominee Brett Kavanaugh and the woman accusing him of sexual assault, Christine Blasey Ford. During the course of the hearing, you can see moments when each was asked a question by the senators and prosecutor, with the colour indicating whether those questions were actually answered or otherwise.

Figure 9.11 Every Time Ford and Kavanaugh Dodged a Question, by Alvin Chang



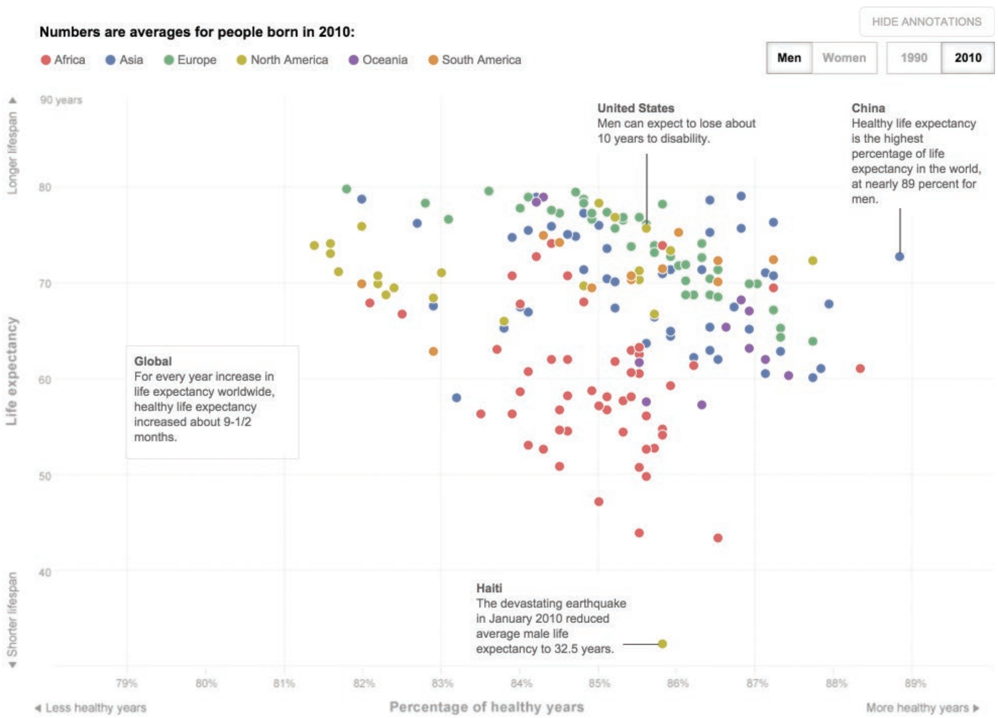


Figure 9.12 How Long Will We Live – And How Well?, by Bonnie Berkowitz, Emily Chow and Todd Lindeman (*Washington Post*)

As the range of different categories grows, you need to expand the range of noticeably different colours. In the scatter plot shown in Figure 9.12, six different hues of colour are used to classify visually all point marks based on which countries are from the different continents of the world.

The ability to preserve clear differentiation becomes harder as the unique colours available diminish. A useful guide to follow is once you exceed 12 categories, there are no longer sufficiently different hues available to assign to categories 13+. There are variations of hues, of course, but they are not different enough to preserve sufficient legibility. Just because variations are available does not make them useful. You will be increasing the viewer's cognitive burden significantly, trying to learn, recall and recognise each association. This delays the process of understanding and undermines the accessibility. There are three ways of handling excessive numbers of distinct categories:

- If interactivity is an option, consider offering filters to modify which category or categories are displayed at any given point in a visualisation, as demonstrated in Chapter 7 in the visualisation of tree species across New York City (Figure 7.3). Alternatively, use a highlighting feature, like the 'Baby names' project example (Figure 7.6) to emphasise some selected values, leaving the remainder presented in grey.
- You might need to loop back to do some further data transformation, by considering how to exclude or combine categories in order to reduce the number of distinct classifying colours needed.

- You may also consider supplementing the use of colour with texture or pattern to create further visible distinctions. In Figure 9.13 you can see two patterns being used occasionally as additive properties to show the structure of tracks on the Beatles' album.

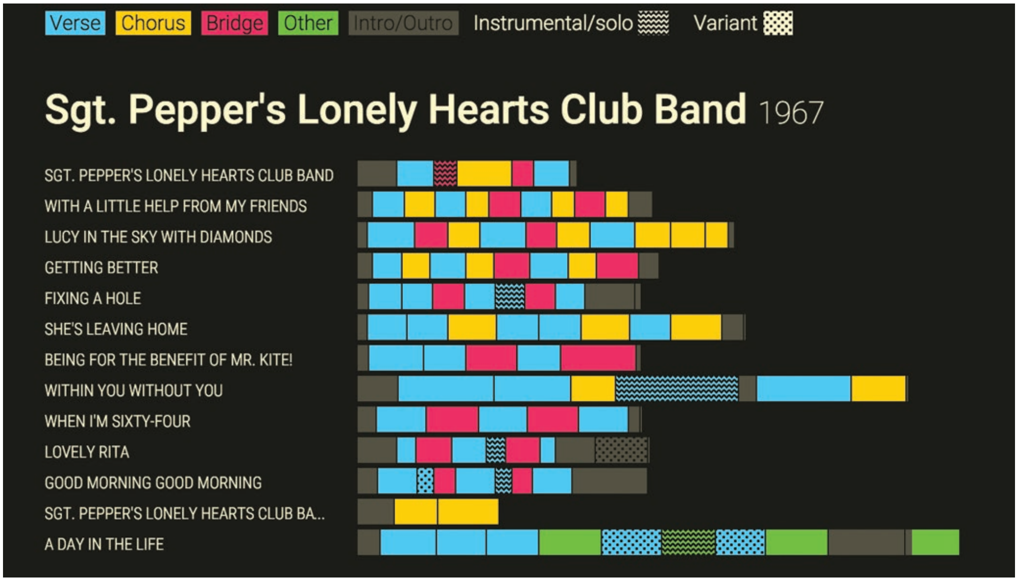


Figure 9.13 Charting the Beatles: Song Structure, by Michael Deal

Sometimes, your categorical data is actually about categories of colour. There can be an immediate explicit relationship between the colours you use and their associated data values. In Figure 9.14, Vienna is reduced to an illustrative 100m2 apartment whereby the floor plan presents the proportional composition of the different types of space and land in the city. The colours and textures of each component explicitly embody the visual characteristics of the associated categorical value. This is another example of acceptable gratuitousness: the colour and appearance demonstrate an additive quality, not a distracting one, creating topic immediacy that accelerates the value recognition.

Another treemap, as shown in Figure 9.15, shows the wide range of different colours used in official rapid transit diagrams of every system in the world. The colours associated with every line on every worldwide metro or subway system are grouped by colour family with the box sizes based on the number of stations served by that line. So, for example, the yellow Circle line is one of 14 different lines on the London Underground system and serves 36 stations along its route.

Ordinal categories are handled a bit differently to nominal categories because they introduce properties of order. When using colour to classify different ordinal categories you are striving not only to create visible distinction between each distinct category, but also to portray the hierarchical relationship that exists between them. The colour approaches used to achieve this tend to align more with how you would represent quantitative values.

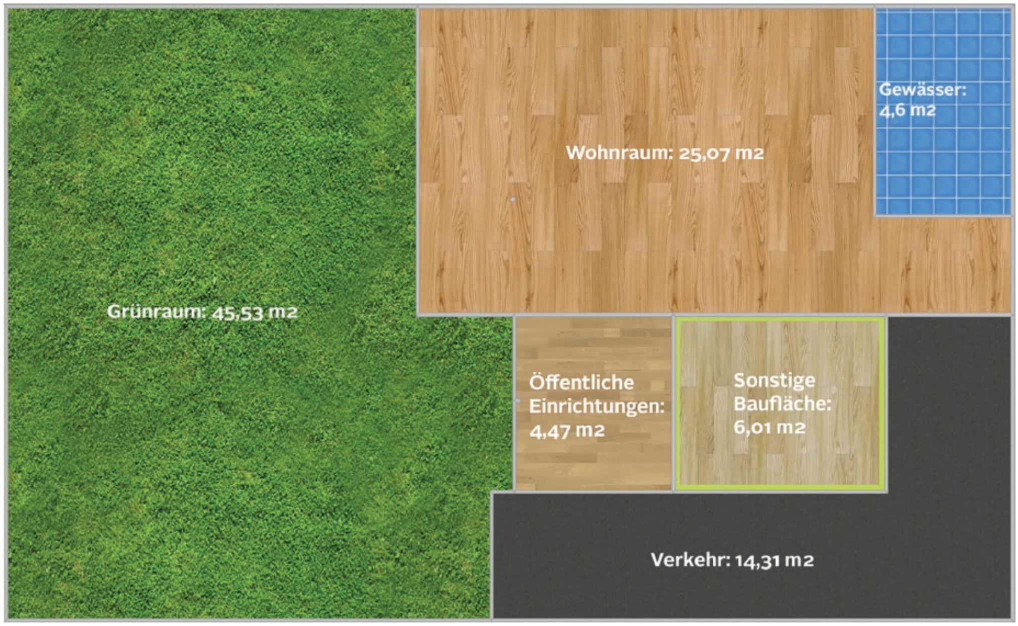


Figure 9.14 If Vienna Would Be an Apartment, by NZZ

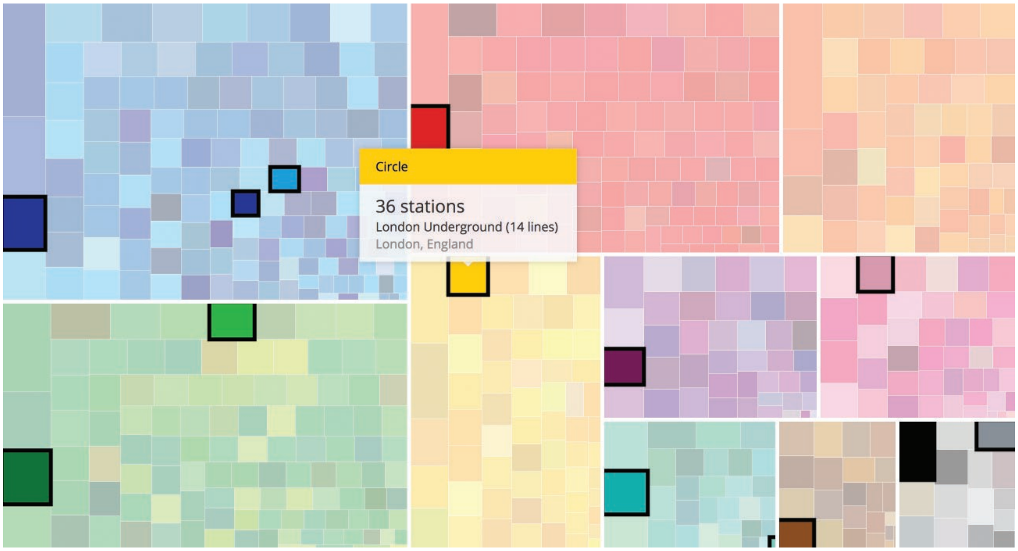


Figure 9.15 Colors of the Rails, by Nicholas Rougeux (www.c82net)

A typical example of a diverging ordinal scale might be seen in a stacked bar chart. The example shown on the left in Figure 9.16 applies ordinal colour classifications to reveal the responses to a range of survey questions. The categories are representative of sentiment and strength of

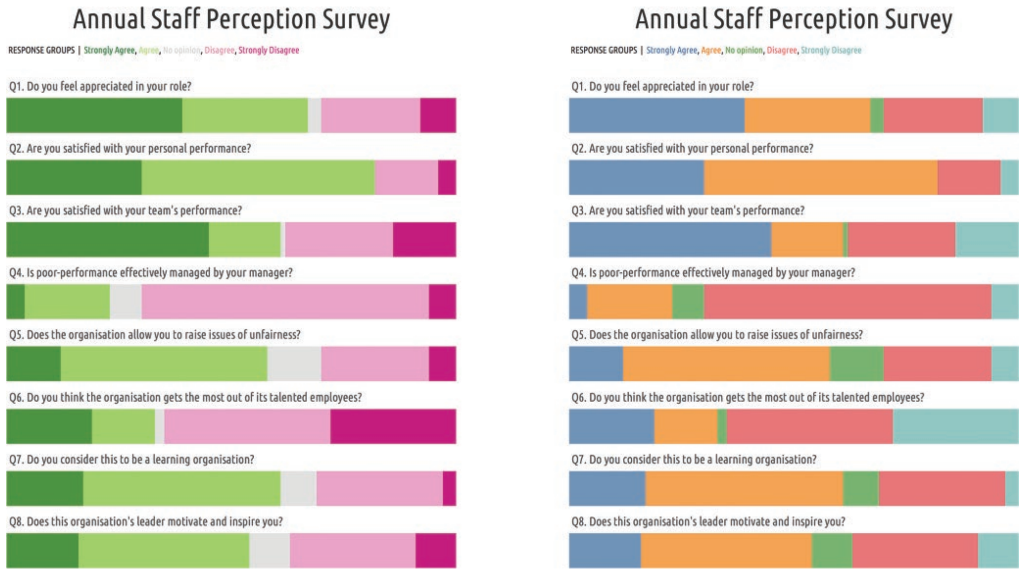


Figure 9.16 Contrasting Approaches to Colouring Stacked Bar Charts Displaying Ordinal Data

feeling, based on a scale from *strongly agree*, *agree*, *no opinion*, *disagree* to *strongly disagree*. By colouring the agreement categories in green, the disagreement categories in pink and 'no opinion' in grey, a viewer can quickly perceive the general balance of feelings being expressed. Darker shades emphasise the strongest feelings at each end of the stacked bar rows.

Although the nominal colouring applied to the same chart on the right-hand side still enables you to learn, look up and see the distinct categories, it fails to make the collective ordinal patterns as discernible as the approach on the left.

Another example of ordinal data might be to represent the notion of recency. In Figure 9.17 you can see a display plotting the 2013 Yosemite National Park fire. Colour is used to display the recorded day-by-day progress of the fire's spread. The colour scale is based on a temporal spectrum with darker shades being *more recent*, lighter tints being *more in the past*. It applies the metaphor of the past having somewhat faded away.

There is an extension in the potential application of ordinal colouring which becomes relevant when you might wish to apply the notion of hierarchical emphasis to draw out significant categorical features of your data that would otherwise merit nominal colouring practices.

This is about drawing contrast between important features that should appear prominently in the foreground for the viewer against other features of less importance that should be more subdued in their appearance. As introduced in Chapter 3, bringing key insights to the surface of your charts contributes towards facilitating an explanatory experience. It bears repeating: if you have something important to say, say it. In this case, say it with colour.

It is here that grey will prove to be a strong ally helping you to convey a sense of depth in your work. You will recall from the opening section that grey is the unsaturated form of a hue. When

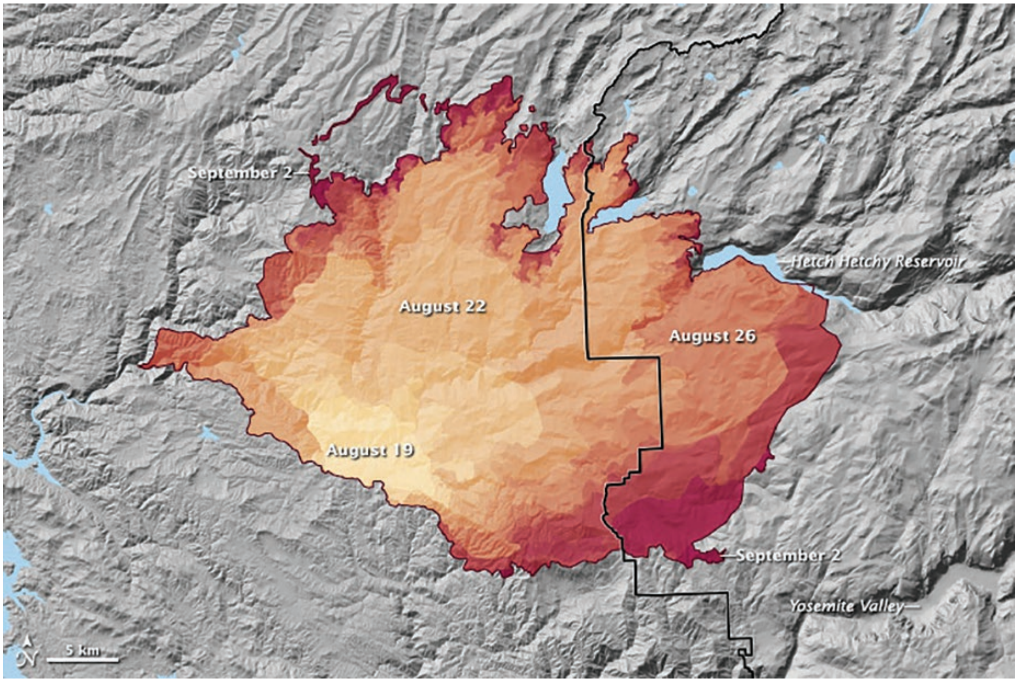


Figure 9.17 'Rim Fire' – The Extent of Fire in the Sierra Nevada Range and Yosemite National Park, 2013: NASA Earth Observatory images by Robert Simmon

an unsaturated colour is juxtaposed with a saturated colour, contrast is achieved. In the bar chart shown in Figure 9.18, the analysis presents a summary of the most prevalent men's names that feature among the CEOs of the S&P 1500 companies. As you can see, there are more guys named 'John' or 'David' than the percentage of *all* the women CEOs combined. With the emphasis of the analysis on this startling statement of inequality, the bar for 'All women' is emphasised in a strong burgundy-coloured hue, contrasting with the grey bars of all the men's names. Notice also that the respective axis and bar value labels are both presented using a bold font for the 'All women' bar, which adds further emphasis.

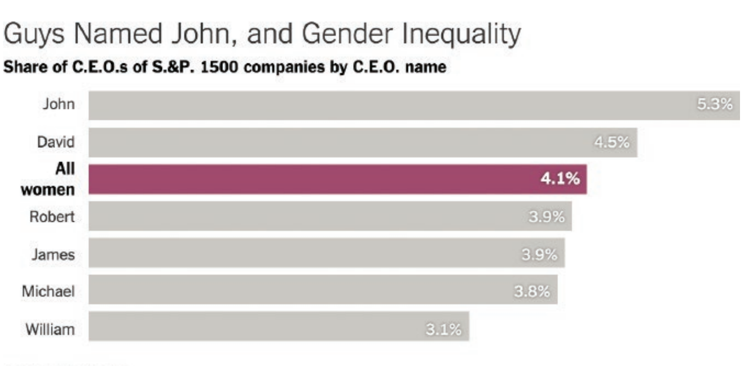


Figure 9.18 Fewer Women Run Big Companies Than Men Named John, by Justin Wolfers (*New York Times*)

Sometimes, only modest emphasis is required. There is no need to shout in order to establish contrast. The chart in Figure 9.19 creates more subtle distinction between the slightly darker shade of green (and emboldened text), emphasising New York's figures, compared with the other listed departments that appear in lighter green. The object of our attention aligns with the topic of interest. In this case, it concerns a drive to recruit more NYPD officers. This does not need to be any more contrasting than it appears; it is sufficiently noticeable and sometimes that is the right level of volume to apply.

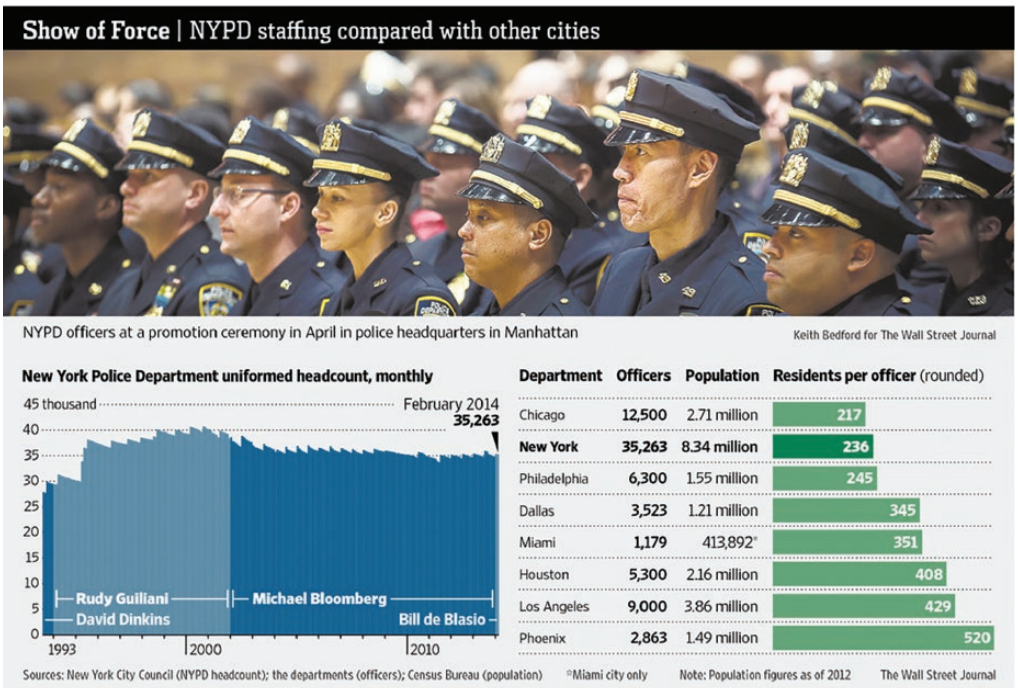


Figure 9.19 NYPD, Council Spar over More Officers, by Graphics Department (*Wall Street Journal*)

The chart shown in Figure 9.20 shows an example of when you want to use colour to establish discrete associations between categorical values, but then also apply an ordinal separation to create contrast between the values you want viewers to read and those included only to provide scale and context. The analysis in this chart looks at international cricketers (specifically batsman) and their cumulative run scoring across their Test match careers based on their age at the time they scored their runs. Seven current or recent celebrated batsmen are elevated to the foreground and categorised using distinct colours for each series line. The rest of the players included in the chart are given a grey shade and thus relegated to the background. We want to see the shapes of their careers, but in this analysis we do not care about finding out who they are.

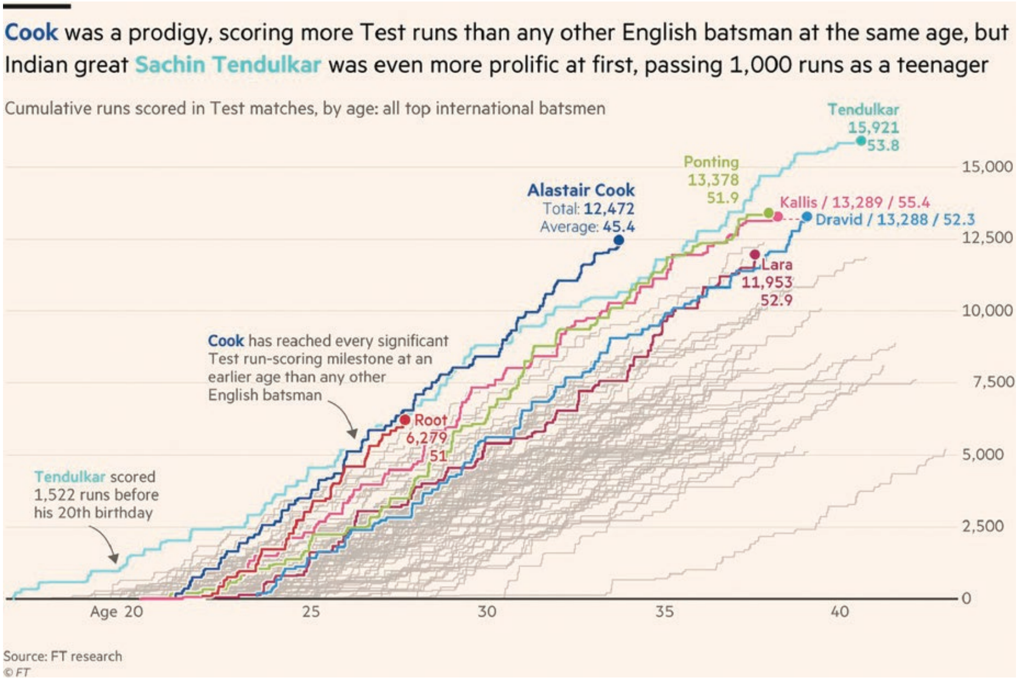


Figure 9.20 Cricketer Alastair Cook Plays His 161st and Final Test Match, by John Burn-Murdoch for Financial Times

Functional Decoration

After making colour choices for optimising data legibility in your charts, you must turn to consider *functional decoration*. This is concerned with colouring every other element of your visualisation display: your interactive features, your chart apparatus and any annotations need to be coloured in order to be visible, but in a way that is harmonious with the colour schemes you have used to represent your data.

Functional decoration may sound like an oxymoron, but it captures the delicate balancing act you face. You have room to experiment in how you might colour your annotations and interactive elements, but only to the extent that those choices do not compromise their functional purpose, which is to support the legibility of data.

There is no single pathway towards achieving this. The Wind Map project (Figure 9.21) conveys a highly aesthetic quality yet uses only a monochromatic palette. There is no colouring of the sea, no topographic detail, no emphasising of any extreme wind speed thresholds being reached. It exhibits artistic *and* functional beauty.

I am not advocating a need to pursue minimalism. Creating something that is pleasing to the eye and equally fit for purpose functionally is a hard balance to achieve. Though you can create elegant work through a limited palette of colours, justifying the use of colours is not the same as unnecessarily restricting the use of colour. Sometimes you will just find a role for

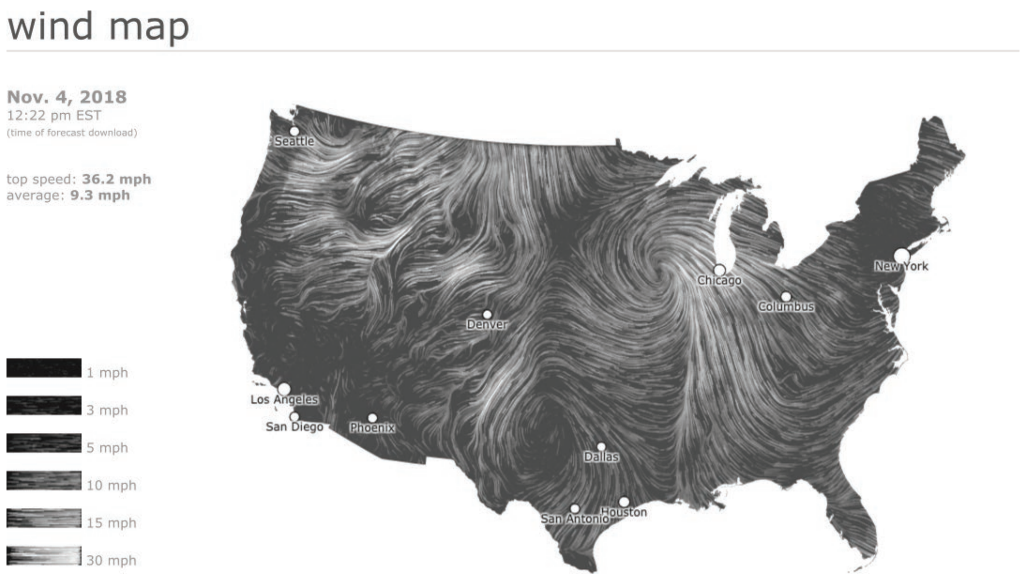


Figure 9.21 Wind Map, by Fernanda Viégas and Martin Wattenberg

many more colours, to help capture the right look and feel for your subject matter. It is why this phase of design thinking is characteristically iterative and often relies on a degree of trial and error in your approach.

Some of the most influential colour practices in data visualisation come from the field of cartography (as do many of the most passionate colour purists). Just think about the amount of visual detail shown in a reference map that relies on colour to help differentiate types of land, indicate the depth of water or the altitude of high ground, mark out routes of road and rail networks, etc. The best maps pack an incredible amount of detail into a single display and yet, somehow, they are still legible and functional.

As with reference maps, every design feature you incorporate into your visualisation display will have a property of colour, otherwise they would be invisible. And all these colour choices are connected. Even though you will often make decisions about colouring features in isolation, there will always be a consequence of that choice elsewhere.

The colour choices for chart annotations, including apparatus like gridlines, axis lines and value labels, are particularly sensitive given their proximity to colours assigned already to the data values. If you choose to classify a category in your data using a shade of grey, using the same grey for your gridlines may create confusion as the eye may lose track of which line relates to which feature.

Additionally, once you commit a colour to mean something you should not use the same colour to mean something different, at least not in the same view or page. Exclusivity in a colour's association is important to preserve for as long as possible so the viewer does not have to relearn its meaning. The graphic on 'Ring-necked Parakeets', featured in Figure 9.22, establishes a quantitative association between a scale of green and pink tints.

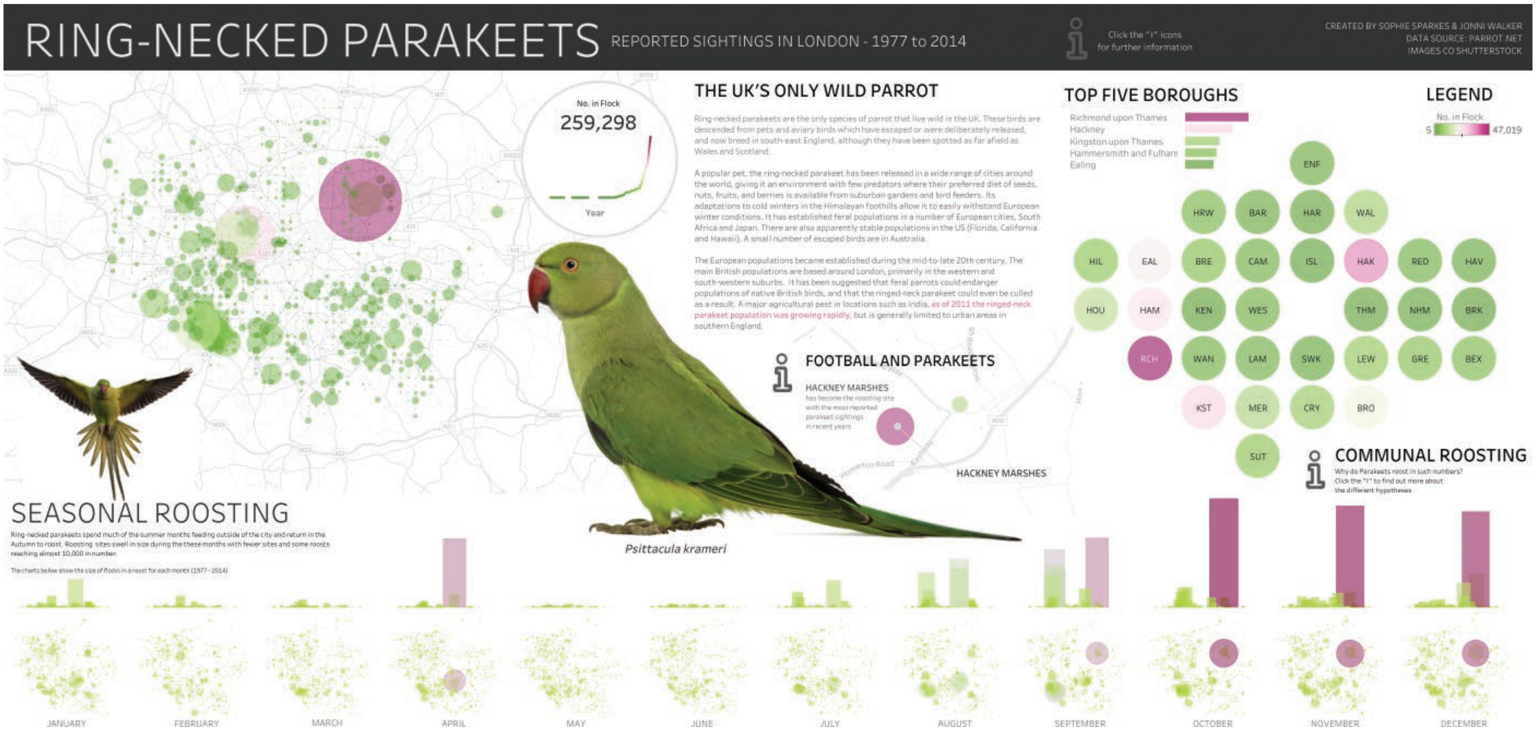


Figure 9.22 Ring-necked Parakeets, by Sophie Sparkes and Jonni Walker

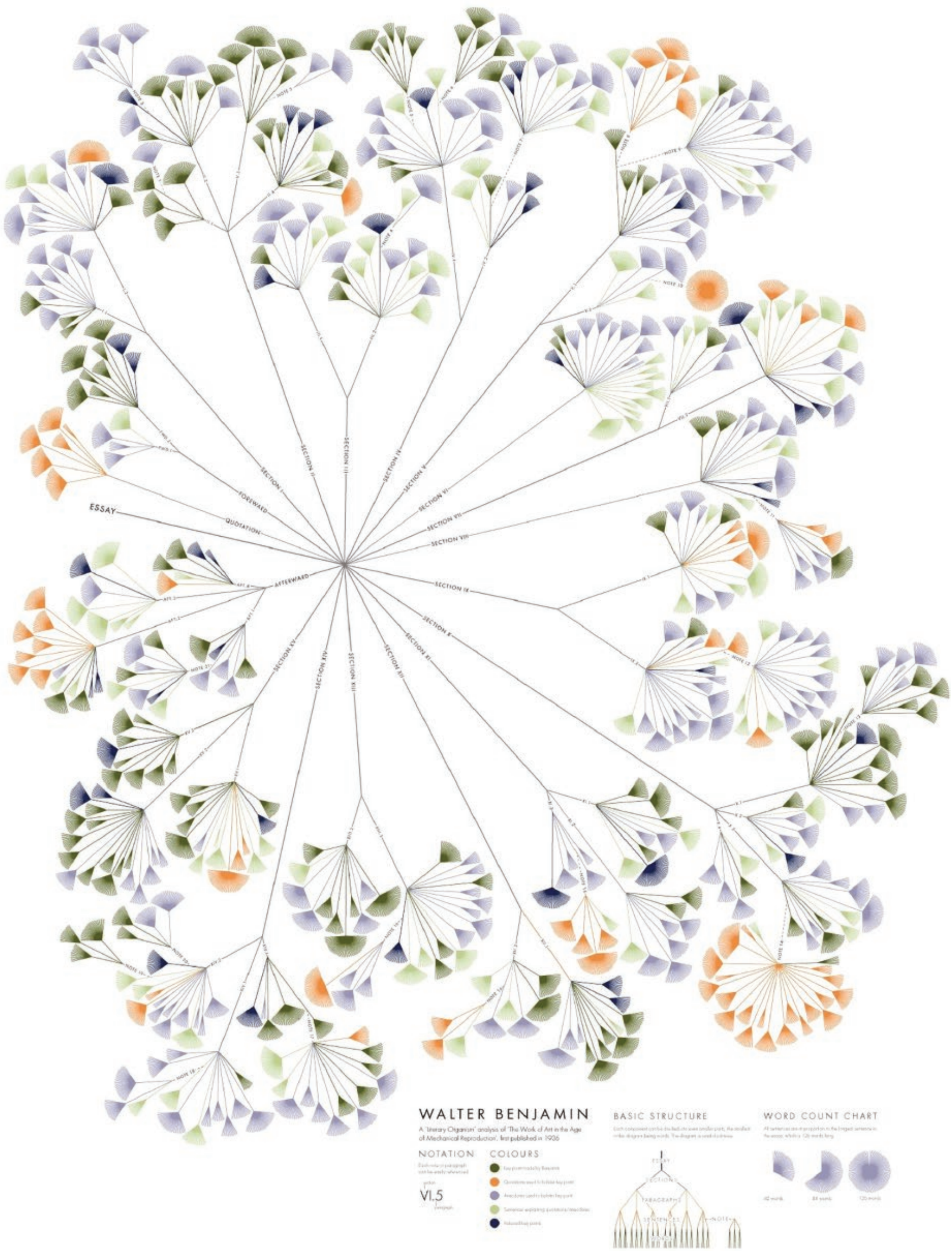


Figure 9.23 Art in the Age of Mechanical Reproduction: Walter Benjamin, by Stefanie Posavec

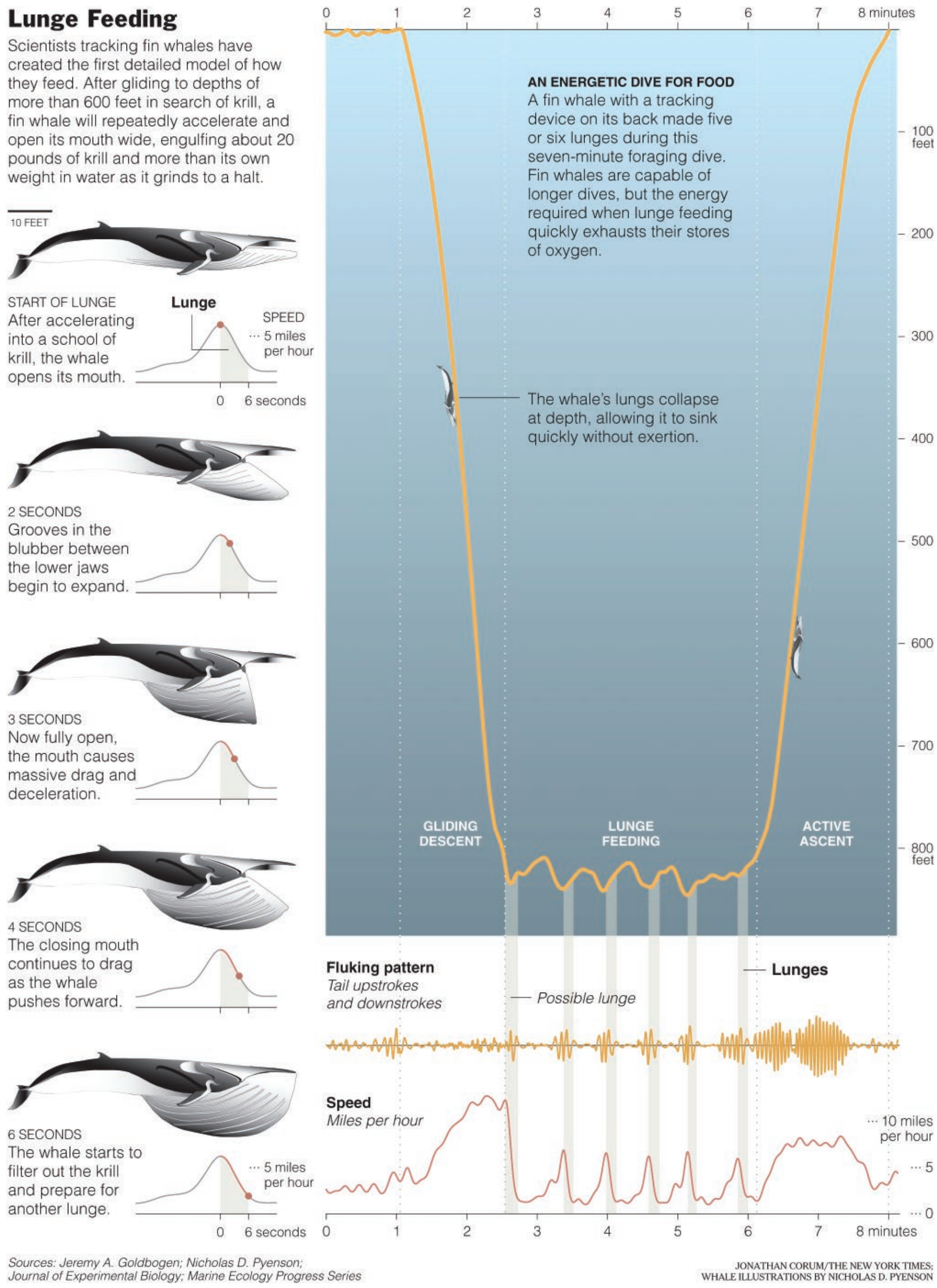


Figure 9.24 Lunge Feeding, by Jonathan Corum (*New York Times*); Whale Illustration by Nicholas D. Pyenson

Once you have learnt this association, you can rely on the same colour associations being continued right across the whole graphic. The viewer can relax into scanning without wondering if the meaning has evolved from one section to another. This significantly amplifies the accessibility of the work and also enhances the elegance through the limited but meaningful colour palette.

If you must use the same colours for different associations, at least try to maximise the 'gap' between each instance of a different association, such as physical gaps (different pages, interactive views), time spans (the duration between reading displays with different associations) or editorial breaks (new subject matter, new angle of analysis). This space will help effectively to cleanse the palate (yes, pun intended) in the mind of the viewer. At each new assignment of a colour, clear explanations are of course mandatory.

> 'When something is not harmonious, it's either boring or chaotic. At one extreme is a visual experience that is so bland that the viewer is not engaged. The human brain will reject under-stimulating information. At the other extreme is a visual experience that is so overdone, so chaotic, that the viewer can't stand to look at it. The human brain rejects what it cannot organise, what it cannot understand.' **Jill Morton, Colour Expert and Researcher**

The quality of harmony across all your colour choices is a hard thing to achieve. It shares the same enigmatic quality as 'elegance', in that you notice it more when it is missing. It is apparent in the colours used by Stefanie Posavec in her visualisation of the structure of Walter Benjamin's essay 'Art in the Age of Mechanical Reproduction' (Figure 9.23). There is an almost effortless cohesion between the colours used across the entire design anatomy of this work: the petals, branches, labels, titles, legend and background.

The 'Lunge Feeding' graphic, Figure 9.24, similarly demonstrates the importance of functional decoration. The blue-shaded panel, getting darker as the sea depth increases down the page, provides a notion of scale for the journey taken by a whale when feeding. This draws contrast from the rest of the layout, establishing the panel as the centrepiece to which all other elements are anchored. The thin grey-shaded columns emerging from the bottom of this panel indicate the occasions of a lunge action, which ties in with the same grey bandings used in the small charts that assist the sequence of illustrations of a whale's feeding action on the left. The style of these illustrations is coherent with the overall tone of the work. Rather than being jarringly different, they feel seamlessly integrated and decorate the work in a functional way, helping the viewer to see the act that is being described.

There are no universal rules for the benefits of any particular colour for shading the background of projects or charts. Your choice will depend on the circumstances and conditions in which your viewers are consuming the work, the inherent association with your subject matter, and the style you are trying to convey. It is not uncommon to see background colours being drawn from the set of neutral options presented in Figure 9.25.

Above all else, the colours you have selected to establish data legibility will be key. In general, a white background gives viewers the best chance of being able to perceive accurately the different colour attributes used in your data encoding and especially scales



Figure 9.25 Examples of Common Background Colour Tones

that use degrees of lightness. In Figure 9.24, the influence relationship was inverted: in using blue to colour the background of the sea, the hue of orange offered the most contrasting option to ensure the path of the whale's dive was most visible. This work also demonstrates the value of emptiness or white space to establish layout. Think of it as visual punctuation, offering moments in your work where the viewer can pause, reflect and then move on to the next discrete element.

With thematic maps, there is often merit in including some kind of reference map in the background to assist with orientation. The dot map in Figure 9.26 looks at the language of tweets posted over a period of time across the New York City area. Given the density and number of discrete data points, permanently and simultaneously including the detailed features shown in the mapping layer becomes too visually cluttered. The developers employ a smart interactive solution to overcome this, offering an adjustable slider that

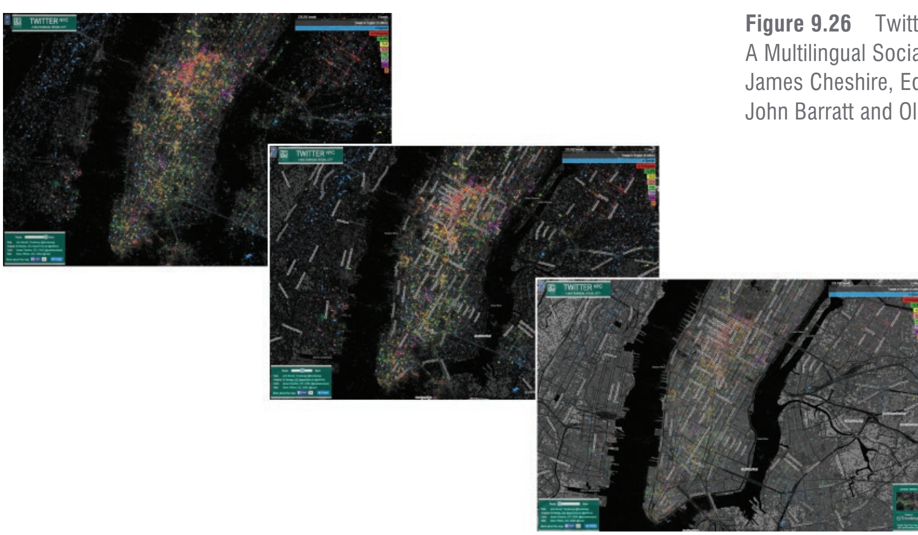


Figure 9.26 Twitter NYC: A Multilingual Social City, by James Cheshire, Ed Manley, John Barratt and Oliver O'Brien

allows users to modify the transparency of the network of roads to reveal the apparatus of the mapping layer.

For interactive projects, every control needs colour in order to be visible, but it must also demonstrate ‘affordance’: properties that indicate what functional events are possible and, where relevant, what events have been activated. The example shown in Figure 9.27 examines the connected stories of casualties and fatalities from the Iraqi and Afghan conflicts. You will see several interactive controls, all of which are astutely coloured in a way that feels consistent with the overall tone of the project, but also makes it functionally evident what each control’s selected setting is. This is achieved through subtle but effective combinations of dark and light greys that help to indicate what has been selected or highlighted, such as the ‘Afghanistan’ and ‘Iraq’ tabs at the top. Filter controls at the bottom use brighter greys to show the selected range of values, but also preserve visibly what other currently unselected values are available to include.

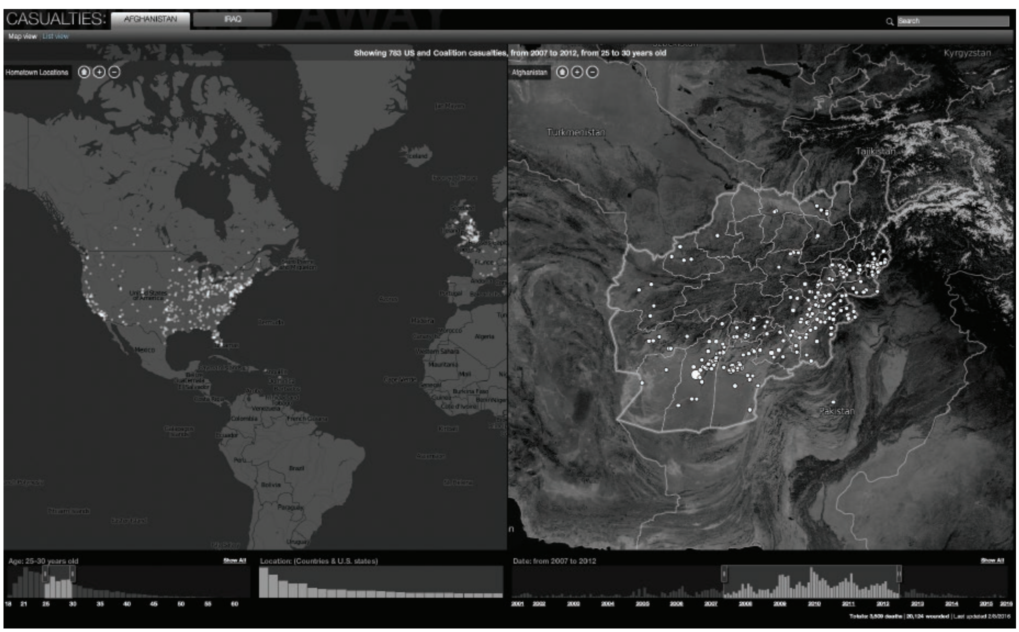


Figure 9.27 Casualties, by Stamen, published by CNN

9.3 Influencing Factors and Considerations

We have already touched on the strong influence of using colour when displaying different types of data, but there are many other factors that will influence your decisions about how colour *should* be used in your work.

Medium: If your printed work will need to be compatible for both colour and black and white output, before finalising your decisions check that the legibility and intended meaning of your colour choices are being maintained across both. It might seem obvious but there is a significant difference in how colours appear when published in *colour* and how they appear when published in *greyscale* because different hues have different levels of brightness. The purest blue is darker than the purest yellow, and so if printed using black and white settings, blue would appear a darker shade of grey and imply it is representative of a higher order of value.

For digital displays, the conditions in which the work will be consumed will have some influence over the choice between, for example, light and dark backgrounds. It can be hard to mitigate for all the subtleties of variation in light present at the time of consumption, but if your work is generally intended for consumption in a light environment, lighter backgrounds with saturated foreground colours tend to be more fitting; likewise, darker backgrounds will work best for consuming in darker settings.

Colour rules: In many organisations, publications and websites, there are style guidelines that require the strict observation of colour rules. These are often established with good intent, driven by a desire to create conformity and consistency in style and appearance. Developing a recognisable 'brand' and not having to think from scratch about what colours to use every time you face a new project are things that can be very helpful, especially across a team environment. However, in my experience, the contents of such colour guides rarely offer optimal application for allocating colours to fulfil a variety of different roles in a data visualisation work. Compromising on the rules would be the ideal scenario, but the main point is to discover the constraints that exist early in your process so you do not arrive at this stage ignorant of the restrictions you will be facing.

Purpose: Colour choices will strongly influence the visible tone of your work. Does it need to be modest or stimulating? Can you select from a vivid and varied palette or should you be striving for more muted and distinguished options? If you are pursuing an explanatory experience, you may have determined that you will be seeking to say something with colour, using it to draw out significant features of your data.

Accessible design: Approximately 5% of the population have visual impairments that compromise their ability to discern particular colours and colour combinations. Deuteranopia is the most common form, often known as red–green colour blindness, and is a particular genetic issue associated with men. The traffic light scheme of green = 'good', red = 'bad' is a common approach for using colour as an indicator. It is a convenient metaphor, especially in the corporate world, and the reasons for its use are entirely understandable. However, as demonstrated in the treemap of Figure 9.28, which has been rendered to simulate deuteranopia, the meaning of reds and greens will not be at all distinguishable and it will be inaccessible to those affected.

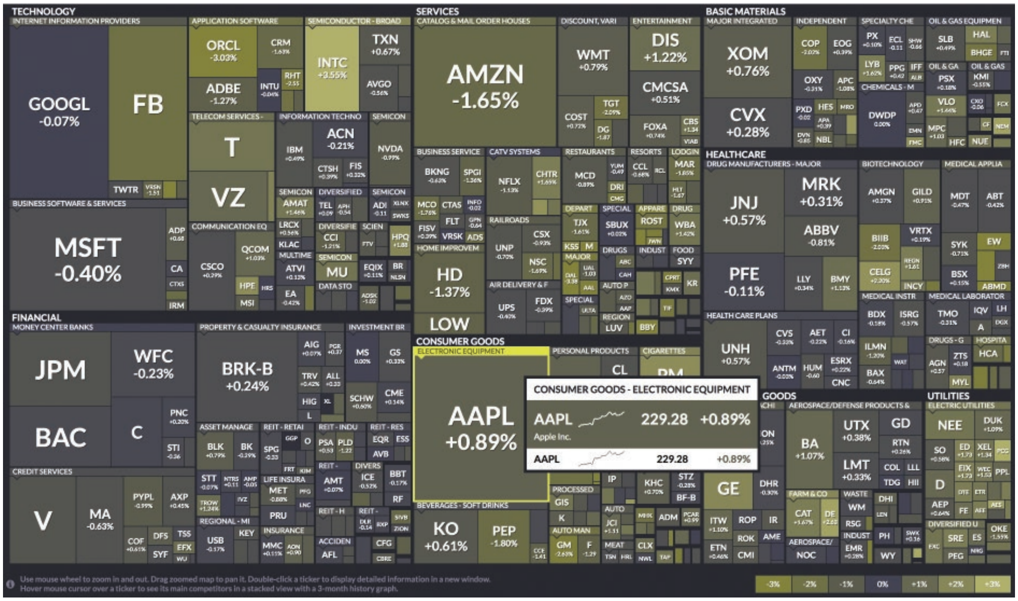


Figure 9.28 Finviz: Standard & Poor's 500 Index Stocks (www.finviz.com)

If nobody in your audience has such visual impairments, it is not necessary to avoid the use of red or green, but if your audience are large and undefined you may need to consider colour-blind-friendly alternatives. Some options are presented in Figure 9.29. The first three options show variations of green tones alongside different secondary pairings that might be considered instead of the standard default red. The fourth option switches the metaphor to red = hot = good, blue = cold = bad. The final option uses secondary encoding through symbols to convey the association if the colours cannot be perceived.



Figure 9.29 Colour-blind-friendly Alternatives to the Standard Green and Red Tones

Deuteranopia is not the only visual impairment to be concerned about. For those with limited eyesight, features that offer magnified views and voice assistance may be worth considering, should they be viable.

An extension of accessible design thinking is to consider the impact of potentially exploiting established colour associations with your subject matter. In politics, sport, brands and in nature,

there are many subjects that already have immediate associations that offer the viewer a shortcut to accelerate recognition. These might be applied to encode your data or functionally decorate your work.

However, while some colours can offer useful and positive associations, in some cases there can be negative connotations that should be handled sensitively or even avoided. You would not use bright, happy colours if you were portraying matters of death or disease. To use a blue colour in a project about depression would be insensitive. Using any notion of skin colour to represent ethnic groups is something that would be understandably considered offensive unless there were very good reasons for skin colour being intrinsic to the data.

Occasionally, established colour associations are out of sync with contemporary culture or society. For example, when you think about colour and the matter of gender, because it has been so endlessly adopted down the years, it is almost impossible not to think instinctively blue for boys, pink for girls. My personal view is that this association should be avoided. I agree with many commentators who say the association of pink to signify the female gender, in particular, is clichéd, outdated and no longer fit for purpose. This is not a universal view, and I have encountered many who disagree with it. However, I do not think it is too much to expect viewers to learn the association of two different colours for representing gender.

Cultural sensitivities and inconsistencies are also important to consider. In China, for example, red is a lucky colour and so the use of red in their stock market displays, for example, indicates rising values. In Western society red is often the signal for a warning or danger.

Summary: Colour

Features of Colour

This chapter introduced you to colour theory and presented different ways of applying colour in a visualisation to facilitate data legibility and deliver functional decoration.

- Data legibility: Using colours to represent different types of data, with distinctions in approach for representing classifications for quantitative data and associations with categorical (nominal) data. A further distinction was made for using colour to emphasise the relationships between ordinal categories.
- Functional decoration: Concerning decisions about applying colour to every other visual element in your work, including interactive features and annotations.

Influencing Factors and Considerations

If these were the options, how did you make your choices? The influencing factors included:

- Medium: The intended output format of your work will affect both colour choices and how they are perceived.
- Colour rules: The need to observe potentially restrictive colour guidelines.

- Purpose: What tone of voice are you trying to convey and how might colour choices shape this?
- Accessible design: Pay attention to potential visual impairments across your audience. Be aware of the sensitivities and positive or negative colour connotations.

General Tips and Tactics

- Use the squint test. Shrink things down and/or half close your eyes to see what coloured properties are most prominent and visible. Are these the right features of your display that should be emphasised?
- Experimentation: Trial and error is often necessary in colour selection, especially for functional decoration.
- Print quality and consistency is a factor. Graphics editors who create work for print newspapers or magazines will often consider using colours as close in tone as possible to pure CMYK, especially if their work is quite intricate in detail. This is because the colour plates used in printing presses will not always be 100% aligned and thus mixtures of colours may be slightly compromised.
- Developing a personal style guide for colour usage saves you having to think from scratch every time and will help your work become more immediately identifiable (which may or may not be an important factor).
- Make life easier by ensuring your preferred (or imposed) colour palettes are loaded up into any tool you are using, even if it is just the tool you are using for analysis rather than for the final presentation of your work.
- If you are creating for print, make sure you do test print runs of the draft work to see how your colours are looking – do not wait for the first print when you (think you) have finished your process. What looks like a perfect colour palette on screen may not ultimately look the same when printed.

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

10

Composition

Composition is the final part of your design anatomy. It concerns the management of space. By definition, composition can be seen as both the *act of* and *result of* arranging a mixture of visual ingredients together to form a final whole.

You should not infer that discussing this topic in the final chapter means it is the least important part of developing your design solution. Far from it. It is just that only after having considered annotation can you reasonably move past thinking about what will be included and then defining how it will appear.

Charts, interactive controls and features of annotations all occupy space. The decisions you make in the final step cover the physical attributes of, and relationships between, every design element that is to be included in your final work. By extension, interactivity can also affect how you use your space, how you overcome the limitations of your space, and how you navigate to other space.

10.1 Features of Composition

Layout

Judging layout is the essence of composition thinking. Well-considered layouts optimise the readability and meaning of the collected content. They are a function of the relative positioning and sizing of all your design elements in the space you are working with. Just as variation in colour implies meaning, so too does variation in position and size. A chart that is larger than another chart will imply it carries greater importance. Charts of equal size but located in different places will lead to extra attention being commanded by the one positioned at the top of a screen or presented first in a sequence.

The purpose of your layout is to establish a hierarchy of meaning and importance, offering the viewers clues about the journey they should take through the content.

In Figure 10.1 we see the infographic 'City of anarchy'. This demonstrates a clear visual hierarchy across its layout. There is a prominent heading that introduces the work and manages, temporarily, to pull your eye away from the temptation of focusing too prematurely on the beautifully illustrated 'cutaway' diagram, which is the centrepiece of the work. Annotated captions orbit around the graphic making interesting observations available when your eye reaches

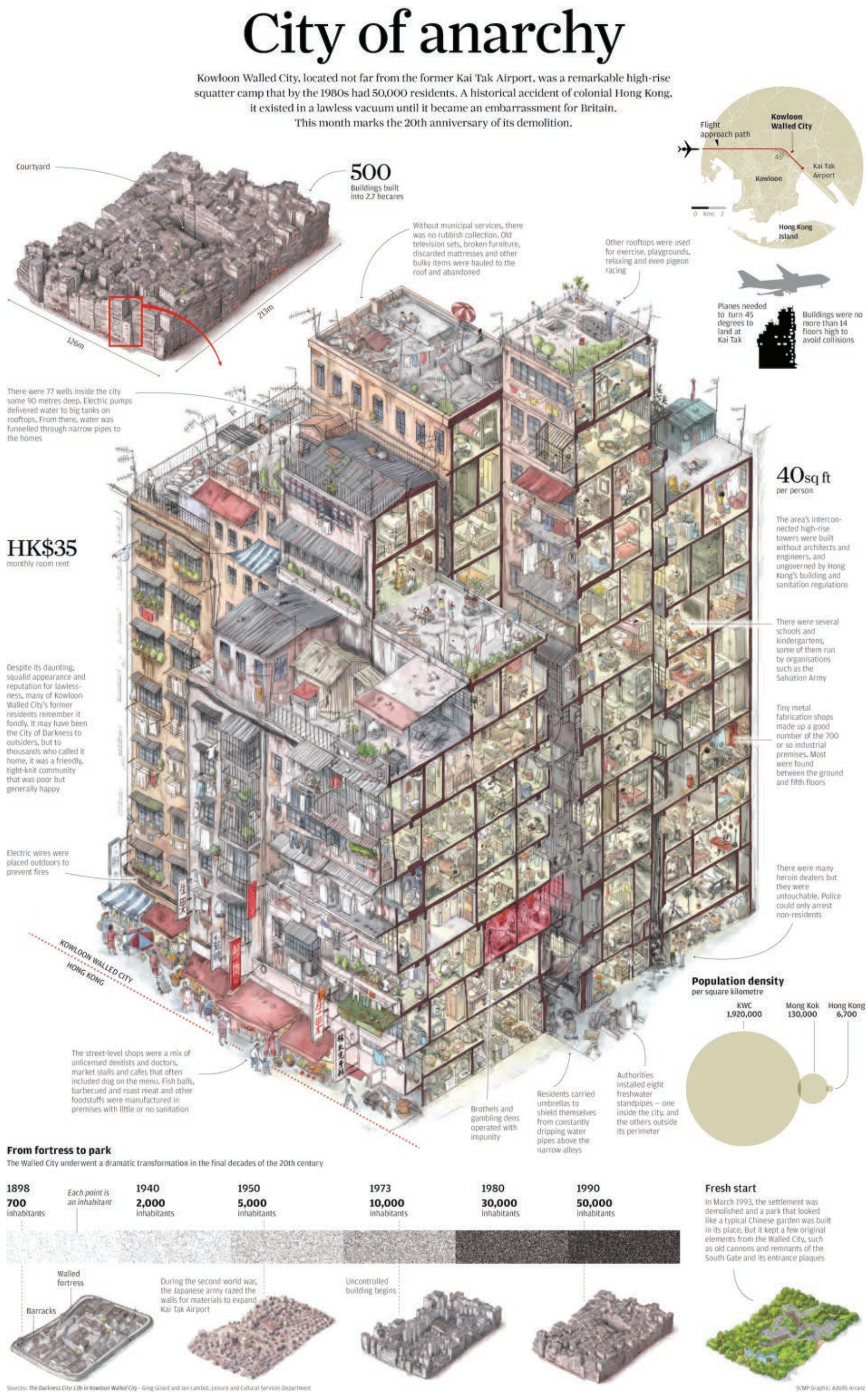


Figure 10.1 City of Anarchy, by Simon Scarr (*South China Morning Post*)

that part of the display. The thumbnail references at the top of the page offer spatial context, providing a localised map view and a world map to explain where this building is located. Those references are there to assist when you want them. At the bottom of the page there are small illustrations to provide supplementary analysis about the history of the city's growth over time. It is clear through their placement at the bottom of the page and their diminutive stature that this analysis will only be encountered much later in the reading process.

When starting to think about your potential layout you cannot usefully isolate your thoughts about position from matters of size. When you arrange furniture in a room the decisions you make about where to put things are informed by how big those things are. But if the absolute size of the furniture is not yet defined then the permutations of different arrangements increase in number substantially.

To break this impasse, there are two approaches to help start shaping your composition ideas: wireframing and storyboarding. *Wireframing* involves creating low-fidelity sketches of the potential layout of all your design elements within a single page, like an infographic or an interactive where all functions apply within the same page or view. Figure 10.2 shows the

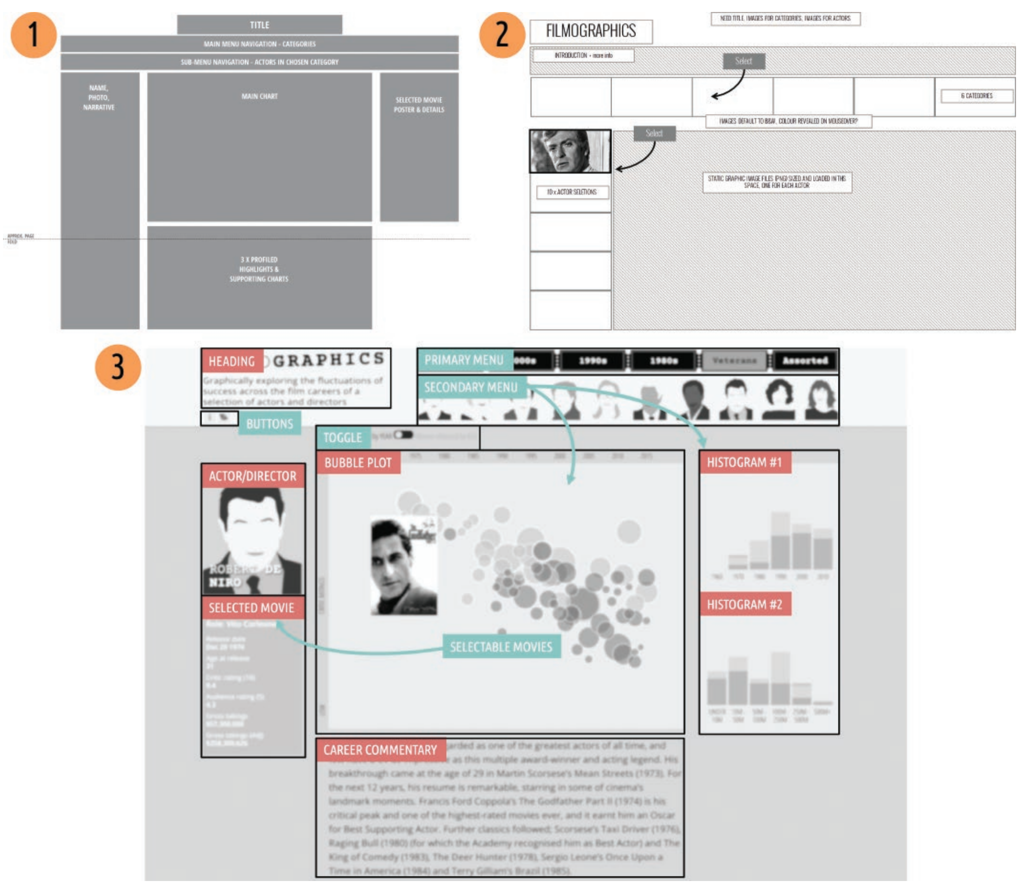


Figure 10.2 Filmographics, by Andy Kirk and Matt Knott

iteration of wireframe designs leading towards the final composition for the 'Filmographics' project. This shows the evolution of ideas for the placement of the charts, the annotations and the interactive controls.

Storyboarding is used to establish the overall structure of your work when it will entail multiple distinct views (e.g. a report or presentation, wide-ranging interactive). It organises your thinking about the sequencing of and navigation between each distinct view of content. The composition within each of these views then goes through more detailed wireframing.

Regardless of whether this activity is carried out using pen and paper, basic tools or more sophisticated technologies, always start with rough ideas and from there the precision will emerge through iteration and experimentation.

It stands to reason that charts will and should be the centrepiece of any visualisation work. Anchoring your layout around where your viewers should encounter your charts can be a useful starting point. Thereafter, the placement of interactive controls and features of annotation should be supporting the experience of understanding, not dominating it.

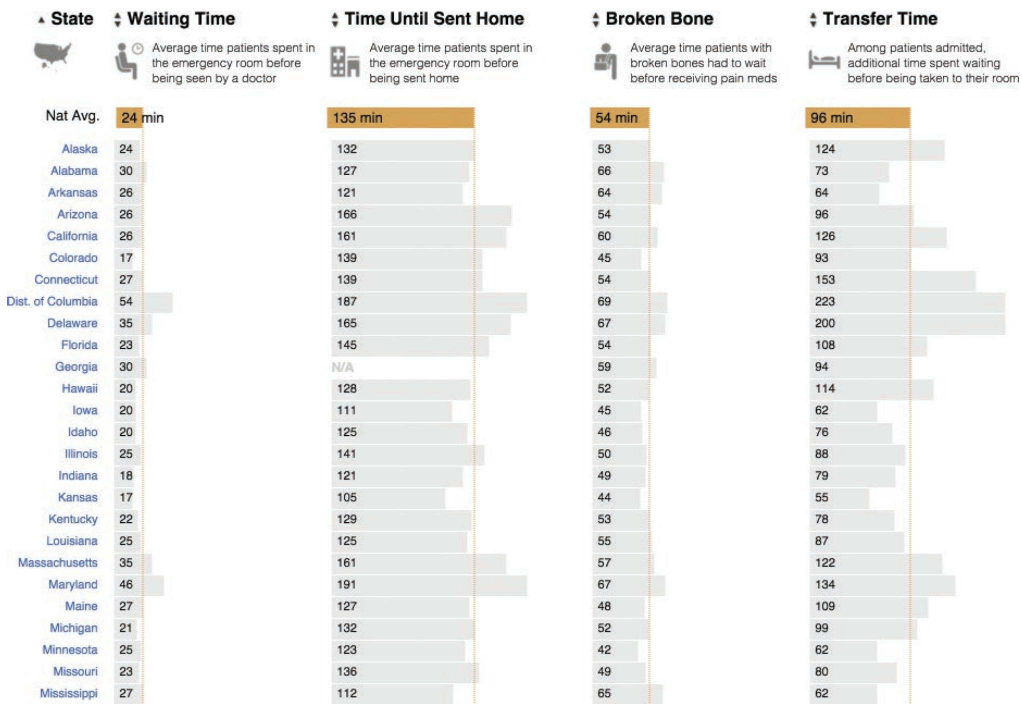
Interactive controls will ideally be located as close to where the functions will be performed so the eye and hand have far less distance to travel between the two. For annotations, the order in which I profiled the different potential features in Chapter 8 was based on the arrangement of where those elements are typically located within a visualisation work. Starting from headings and introductions, moving through reader and/or user guides, on to chart-related apparatus and labelling, and then finishing with footnotes and methods. You might need the setup of introductions and guides to be seen before any charts are seen in order to enhance the reading process. Although there are certain conventions you might follow for the efficiency of your thought process, you have the freedom to determine whatever is the best layout suitable for your project's purpose.

Arranging

Arranging concerns the ordering and direction of your data content as it is displayed within a chart. This is an important consideration to help viewers perceive your representations in the most relevant formation. There are several different approaches to sorting data.

Alphabetical sorting is a cataloguing approach that facilitates efficient lookup and reference of textual or categorical values. You would use this arrangement when you need to offer your viewers an efficient way to look up specific categorical values when there are many items included. In Figure 10.3, investigating different measures of waiting times in emergency rooms across the USA, the bar charts are presented using alphabetical sorting of each state name. This is the default setting, but users can also choose to reorder the bars in other ways across the other columns.

Alphabetical sorting is a common approach but one that often reflects a default choice rather than a useful one. It can be the least interesting way to arrange your data values. However, it might be seen as a suitably diplomatic option should you need to avoid politically displaying your data using any form of ranking, in particular. Additionally, it is absolutely sensible to employ alphabetical ordering for values listed in interactive controls like dropdown menus.



State	Waiting Time	Time Until Sent Home	Broken Bone	Transfer Time
	Average time patients spent in the emergency room before being seen by a doctor	Average time patients spent in the emergency room before being sent home	Average time patients with broken bones had to wait before receiving pain meds	Among patients admitted, additional time spent waiting before being taken to their room
Nat Avg.	24 min	135 min	54 min	96 min
Alaska	24	132	53	124
Alabama	30	127	66	73
Arkansas	26	121	64	64
Arizona	26	166	54	96
California	26	161	60	126
Colorado	17	139	45	93
Connecticut	27	139	54	153
Dist. of Columbia	54	187	69	223
Delaware	35	165	67	200
Florida	23	145	54	108
Georgia	30	N/A	59	94
Hawaii	20	128	52	114
Iowa	20	111	45	62
Idaho	20	125	46	76
Illinois	25	141	50	88
Indiana	18	121	49	79
Kansas	17	105	44	55
Kentucky	22	129	53	78
Louisiana	25	125	55	87
Massachusetts	35	161	57	122
Maryland	46	191	67	134
Maine	27	127	48	109
Michigan	21	132	52	99
Minnesota	25	123	42	62
Missouri	23	136	49	80
Mississippi	27	112	65	62

Figure 10.3 ER Wait Watcher: Which Emergency Room Will See You the Fastest?, by Lena Groeger, Mike Tigas and Sisi Wei (ProPublica)

Chronological sorting is used when the data has a temporal dimension and you need your display to expose patterns over time. In Figure 10.4, you can see a snapshot of a graphic that portrays the rain patterns in Hong Kong since 1990. Each row of data represents a full year of daily readings running from left to right, with the years arranged vertically from the past to the present.

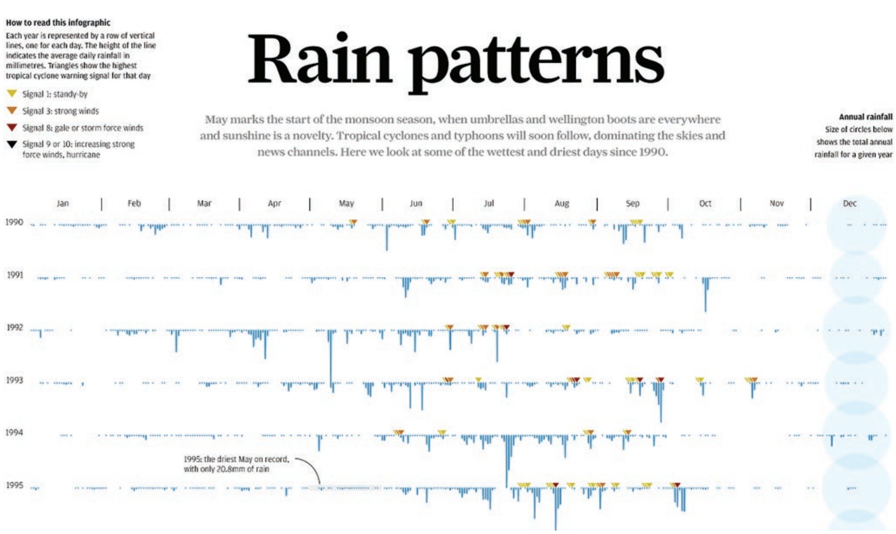


Figure 10.4 Extract from Rain Patterns, by Jane Pong (*South China Morning Post*)

Locational sorting involves sequencing content according to a spatial dimension, particularly when you are not using a mapping technique to portray your data and it is more about relative, rather than fixed, location relationships. This could involve sorting values based on geographic relationships (such as presenting data for all the stations along a train route) or a non-geographic relationship (like a sequence of values based on the positions of parts of the body, from head to toe). Ordering data by location will only be relevant if you believe there is interest in or significance between comparing adjacent locations. An example of this approach is exhibited by the project 'On Broadway' (Figure 10.5), which is an interactive installation that stitched together a bunch of different data measurement and media items relating to intervals of life along Broadway. This collective work offers compelling views of the fluctuating characteristics of the different communities and neighbourhoods as you journey down the spine of New York City, stretching 13 miles (21km) across and beyond the length of Manhattan.

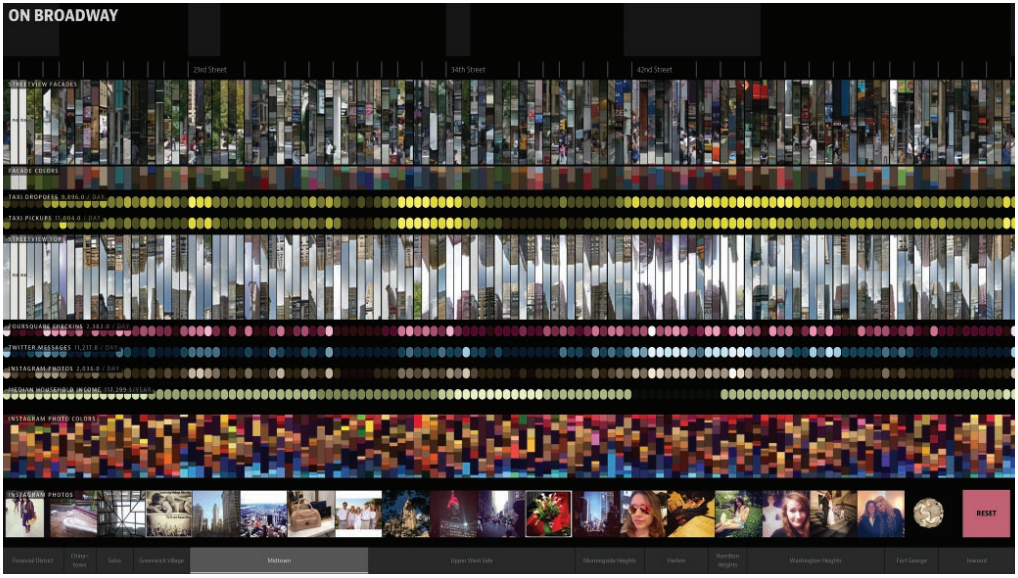


Figure 10.5 On Broadway, by Daniel Goddemeyer, Moritz Stefaner, Dominikus Baur and Lev Manovich

Ordinal sorting can be usefully applied to arrange categorical data that has characteristics of ordered and, potentially, hierarchical relationships. In Figure 10.6, you can see a dendrogram that looks at the consequences of two major beer brands merging, namely SAB Miller and ABInBev. The diagram shows all the individual beer brands that were previously owned by each discrete brand. The hierarchical layout organises this display around a radial structure, starting from the inside tier of a continent, moving out to countries, and then finally to the outer nodes detailing each product.

Finally, *ranked* sorting may be the most common and useful way to arrange data values based on ascending or descending quantitative rankings. In Figure 10.7, the bar chart shows the artists and songs that have held the number one position in the UK charts for the greatest number of weeks. The values are arranged in descending order from the longest duration to the shortest.

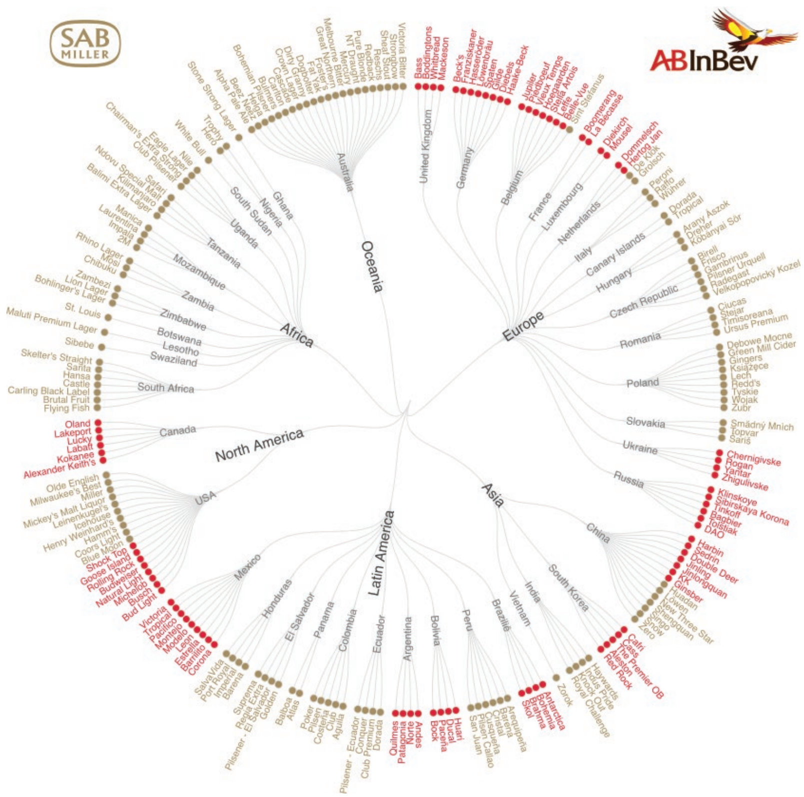


Figure 10.6 The 200+ Beer Brands of SAB and AB InBev, by Maarten Lambrechts for Mediafin

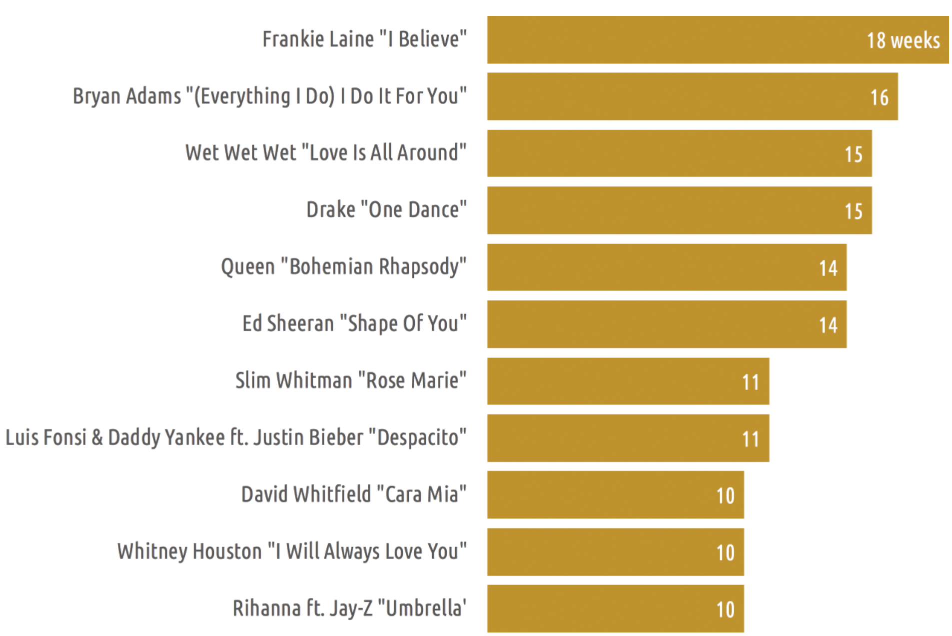


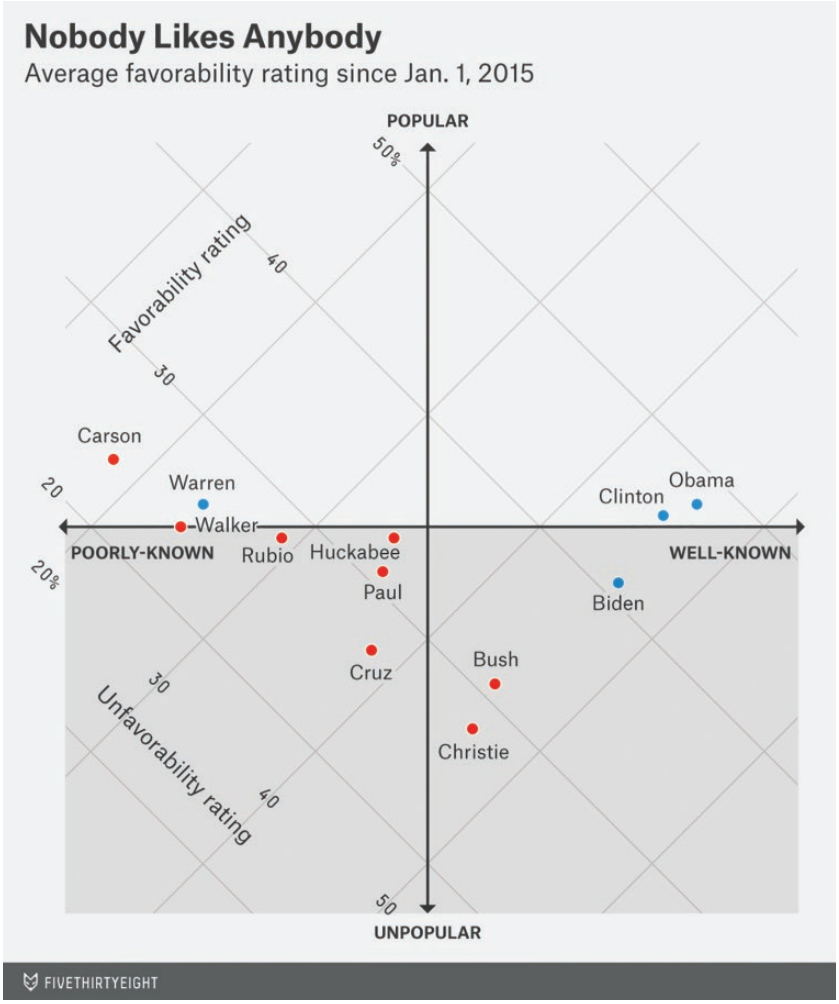
Figure 10.7 The Songs That Were #1 in the UK Charts for the Greatest Number of Weeks

There are further arrangement decisions to make with certain chart types that have overlapping lines, like line charts or bump charts, or criss-crossing connection bandings, like Sankey diagrams and chord diagrams. These charts introduce the need to contemplate value sorting in a z-dimension: that is, which of these features should be displayed on top and which underneath, and why?

The orientation of your chart is another arrangement consideration which might help squeeze out an extra degree of readability and meaning from your display.

The primary thought about chart orientation regards the readability of axis value labels. A vertically arranged bar chart, with multiple categories along the x-axis, will potentially cause viewers to tilt their neck in order to read the labels. You could try adjusting the orientation of the labels to 45° or 90°, but my preference is to transpose the chart so the labels are on the

Figure 10.8 Kasich Could Be the GOP's Moderate Backstop, by FiveThirtyEight



y-axis, the bar sizes are directed along the x-axis, and the category labels are presented in a more readable fashion.

The example in Figure 10.8 rotates a scatter plot by 45° to help guide the viewer's interpretation of what it means for a point mark to be found in each quadrant region. It is also used to emphasise the distinction between being in the top half and bottom half of the chart, which defines the degree of popularity, as this is the principal angle of interest.

Sizing

Do not be afraid to shrink your charts. The eye can detect, with great efficiency even at small resolution, variation in size, position, colour and pattern. The technique of 'small multiples' is commonly used to replicate distinct chart displays for multiple categories or points in time and arrange them usually in a grid layout. This enables the eye to compare and contrast features across all charts in a simultaneous view. Otherwise, you might have to browse through multiple pages or navigate through different selections using interactivity and remember each chart view in order to compare against it. The main obstacle to shrinking chart displays is the impact on text size. The eye will not cope too well with small fonts for axis and value labels, so there has to be a trade-off.

The project featured in Figure 10.9 demonstrates a beautiful example of small multiples. This work is called 'Coral Cities' and looks at how easy it is for people to move within and out of cities. The organic forms displayed represent the distance and routes that can be reached within 30 minutes by car when leaving each city centre. The 40 cities selected are based on the Mercer 'Quality of Living City' rankings. Although created for a large print output, even when looking at a smaller scaled version, any viewer can investigate the shape of patterns for each city but also seamlessly turn their attention to look at all the patterns collectively to explore and find commonalities and exceptions.

> 'Using our eyes to switch between different views that are visible simultaneously has much lower cognitive load than consulting our memory to compare a current view with what was seen before.' **Professor Tamara Munzner, Department of Computer Science, University of British Columbia**

Decisions about chart sizing extend to defining axis scales and value intervals. Although the clues about the most meaningful range of values to include will be shaped by your work at the data examination stage of the process, there are certain conventions that also need to be observed.

When a chart encodes quantitative values using size, the viewer needs to see the full, representative size of the mark, otherwise it will be a distortion of the truth. For example, with bar charts that show data from a common baseline, viewers need to see the full size of a bar based on the value it represents, nothing more and nothing less. To do this you must

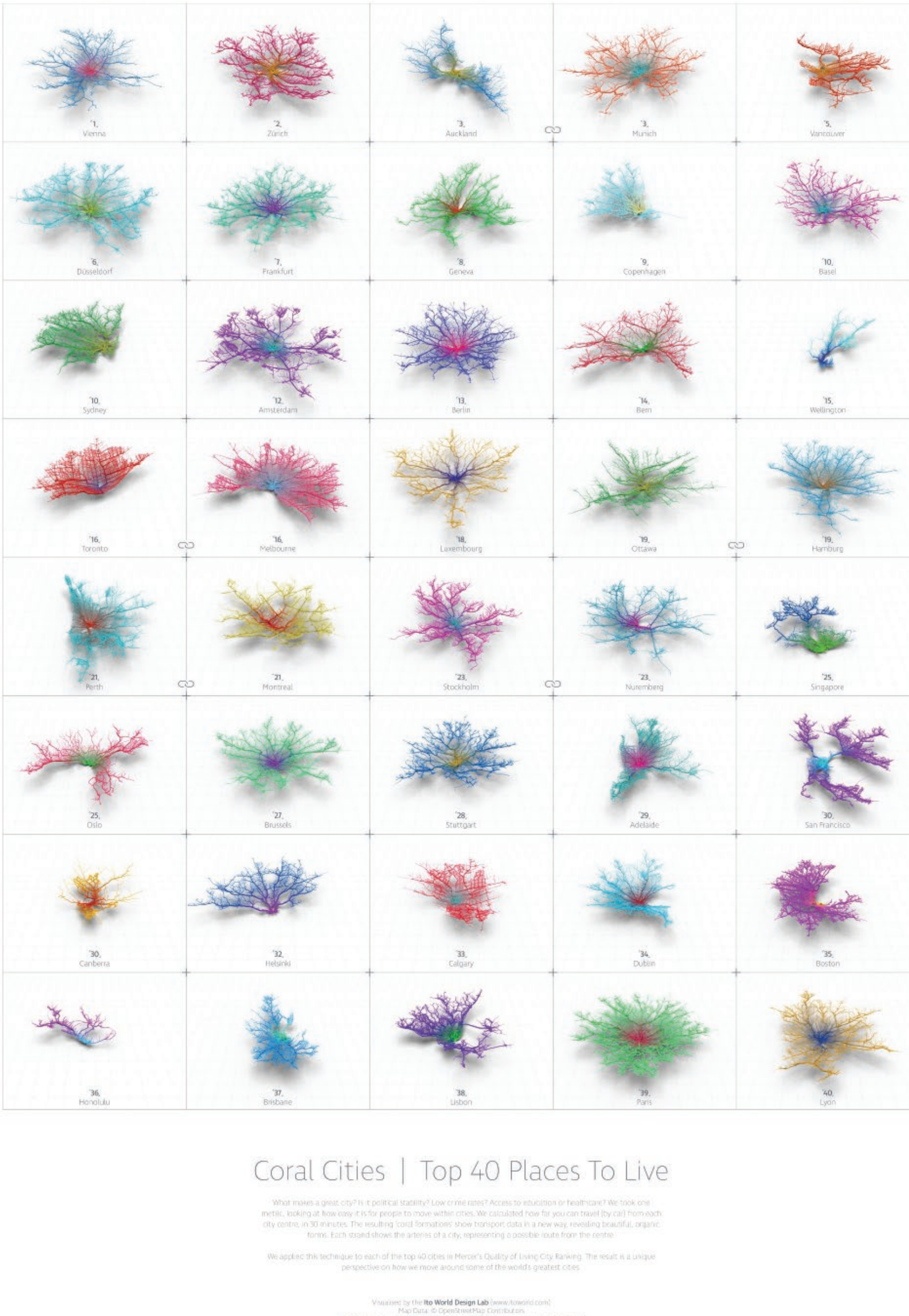


Figure 10.9 Coral Cities, by Craig Taylor, Data Visualisation Design Manager at Ito World

set the origin of the quantitative value axis to zero. If you start this baseline position from any other value, the effect will be to truncate the axis range and the perceived size of the bars. This creates a distortion: the viewer is only presented with part of a bar's true size. The charts in Figure 10.10 show the tallest buildings and structures around the world that are at least 250m in height. The scale used in the chart with the red bars uses a quantitative axis with an origin of 250. This distorts the lengths compared with the true sizing as shown in the first bar chart with the blue bars, which has an origin of 0.

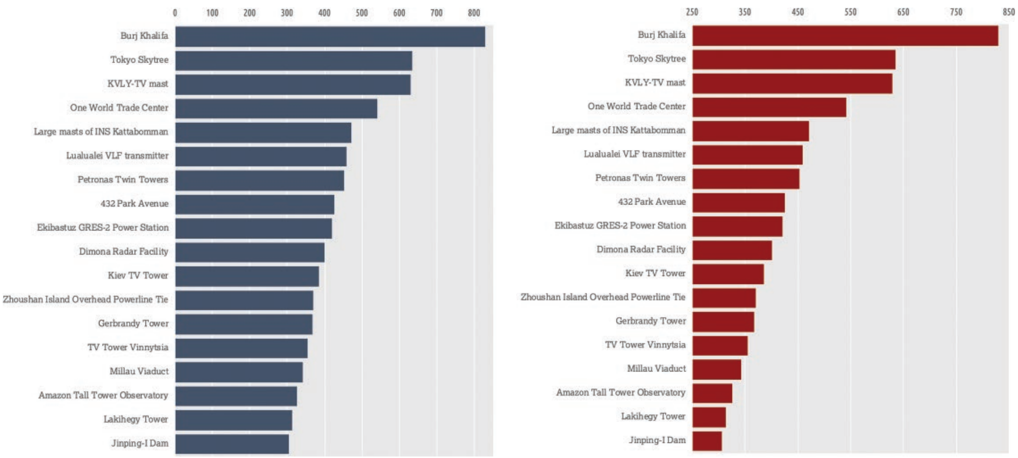


Figure 10.10 Illustrating the Effect of Truncating Quantitative Axis Scales for Bar Charts

Not all charts that use bars necessarily need to start from a zero origin. Variations in the use of the waterfall chart, for example, might be used to show quantitative changes or differences between absolute values (the 'delta'). In this case the base of a bar may be positioned to start from any quantitative position and not necessarily from zero. So long as you still encode its full representative size, that is fine.

In contrast to the bar chart, a line chart does not necessarily need to have a value axis origin always set to zero. It encodes quantitative values through point marks positioned along a value scale, not through size. Truncating the quantitative value axis may be relevant when the notion of a quantity of zero might represent an impossible measurement. It should be made clear to the viewer through clear axis labelling that the baseline position does not represent zero.

In the chart shown in Figure 10.11, we see a line chart plotting the history of 100m record times. Although results have quickened, there is a physical limit to what humans can achieve: running the 100m in a time that is anywhere near zero seconds is impossible. So, starting the value axis at 9 seconds through to a maximum of 11 seconds provides a reasonable axis range in which to plot the observed measurements.

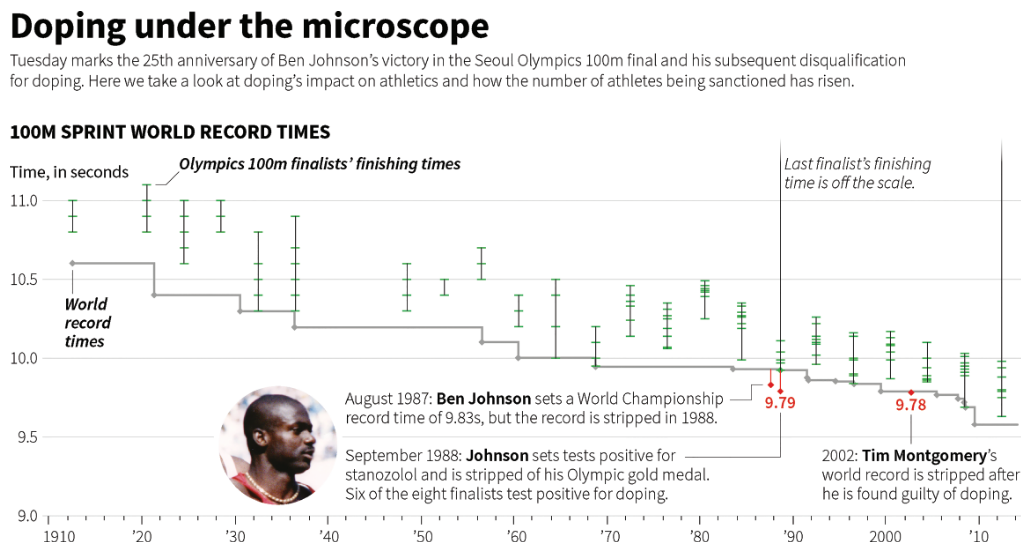


Figure 10.11 Doping under the Microscope, by S. Scarr and W. Foo (Reuters Graphics)

10.2 Influencing Factors and Considerations

Decision making about composition is greatly shaped by common sense but equally burdened by the unsatisfactory shrug of 'it depends'. Here are some of the main considerations.

Medium: Naturally, as composition is about spatial arrangement, the nature and dimensions of the canvas you have to work with will have a fundamental bearing on the decisions you make. There are two concerns here: what will be the shape and size of the primary format; and how transferable will your solution be across the different platforms on which it might be used or consumed?

Decisions about layout will vary depending on whether your work is to be published on a single-page or screen view, such as an infographic or interactive visualisation, or across multiple linear pages (like a report, presentation) or multiple different views (interactively driven navigation controlled by the user). The best solutions composition-wise will vary considerably for each.

The varied characteristics of modern devices present visualisers (or perhaps more appropriately, at this stage, developers) with real challenges. Getting a visualisation to work consistently, flexibly and portably across device types, browsers and screen dimensions (smartphone, tablet, desktop) is hard.

Although many organisations, especially the media, are focused on a mobile-first strategy, the reduced canvas size and the intricacies of requiring users to interact precisely with diminutive controls create difficulties. Given the choice, most visualisers would probably see the desktop as their preferred canvas. Solutions designed for mobile and, to a certain extent, tablet will aim to preserve as much continuity in the core experience as possible but may require certain compromises.

For ProPublica's work on 'Losing Ground' (Figure 10.12), the approach taken to determine what degree of cross-platform compatibility should be preserved was informed by using the heuristic 'smallify or simplify'. Features that worked on ProPublica's primary platform of the desktop would either be simplified to function practically on mobile or just be reduced in size. You will see in the pair of contrasting images how the map display is both shrunk and cropped, and the introductory text is stripped back to include only the most essential information.



Figure 10.12 Losing Ground, by Bob Marshall, The Lens, Brian Jacobs and Al Shaw (ProPublica)

Another consideration about medium will relate to whether your work will be published as an interactive for the Web and also as a static piece for print. The features that make up an effective interactive project may not necessarily translate directly into static form. You might need to pursue two parallel solutions to suit the respective characteristics of each output format.

Quantitative value range: When discussing the physical properties of data in Chapter 4, I described the influence of the shape of your data on your chart composition choices. If you have 30 distinct categorical values in your data, and they all need to be shown, you will need to allocate space for 30 categorical items in your chart layout. The lengths of the words of each item will also need to fit as labels in or adjacent to the chart.

We have already discussed the conventions of setting axis scales, but there are particular composition challenges when you have a wide range of quantitative values. Legitimate outliers will potentially distort your ideal scale choices but will need to be accommodated somehow in the space you are working with.

One solution for dealing with this is to use a non-linear logarithmic (often just known as a 'log') scale. Essentially, each major interval along a log scale increases the value at that marked position by a factor of 10 (or by one order of magnitude) rather than by equal increments.

In Figure 10.13, looking at ratings for thousands of different board games, the x-axis is presented on a log scale in order to accommodate the wide range of values for the 'Number of ratings' measure. This also helps to fit the chart into a neat square layout, which can sometimes be a requirement to enable graphics to be optimally sized for publishing on social media platforms. Had this x-axis remained as a linear scale, in order to preserve this square layout the values below 1000 would have had to be squashed into such a tightly packed space that you would hardly see the patterns. A wide, rectangular chart would have been necessary but impractical, given the limitations of the space this chart would occupy.

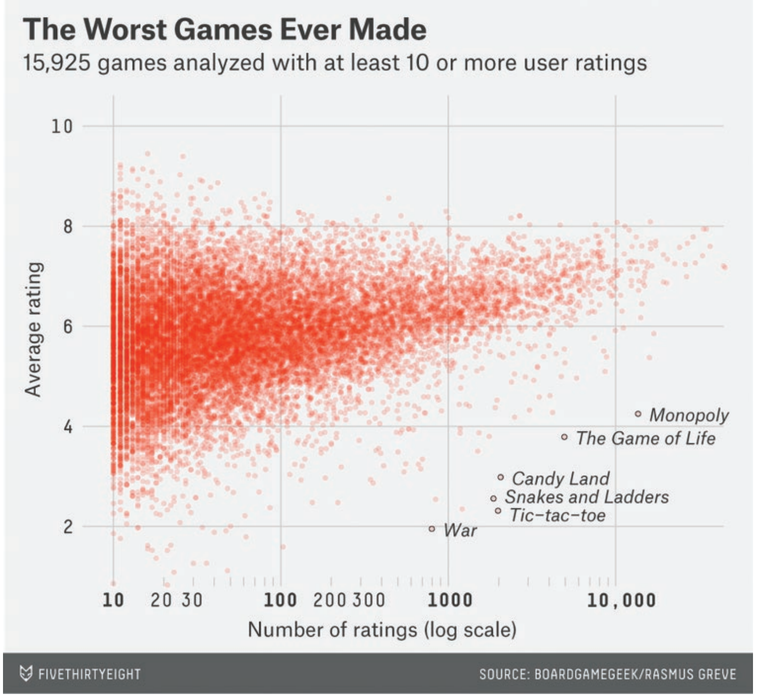


Figure 10.13 The Worst Board Games Ever Invented, by FiveThirtyEight

Editorial thinking: Your editorial thoughts will probably have extended to establishing a sense of hierarchy that might inform which analysis should be displayed more prominently and which less so.

If you have decided to include a number of different angles of analysis in your work, this will amplify the challenge of composition. The more analysis you include naturally increases the demands on space. You might need to compromise by reducing the size of your chart elements or by offering a non-simultaneous arrangement, through multi-page layouts or sequences reached through interactive navigation.

Data representation: Different charts bring different spatial consequences. A treemap generally occupies far more space than a pie chart simply because there are usually many more 'parts' being shown. A polar chart is circular in shape, whereas a waffle chart is square.

Each chart you include introduces a uniquely shaped element that will need to be arranged into your layout.

The table in Figure 10.14 summarises the main chart structures and the typical shapes they occupy, based on the chart types profiled in Chapter 6.

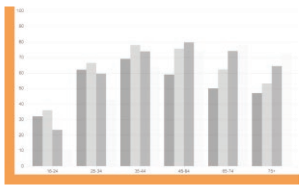
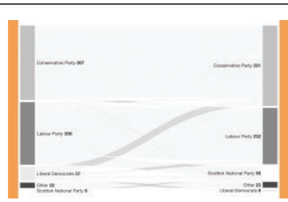
STRUCTURE	SHAPE	DESCRIPTION
Cartesian		These are effectively rectangular structures based on a coordinate system with magnitudes or positions along an x (horizontal) and y (vertical) dimension. The bar and line charts use this structure.
Enclosure		Enclosure charts are based around a fixed shaped container within which data is arranged optimally. This would be seen in the treemap and the waffle chart.
Radial		Radial structures are characterised by a centralised or circular layout usually based on the division of angular parts or axes radiating outwards. They are used for polar and pie charts. Certain hierarchical and relational charts also demonstrate a similar graphical structure, whereby concentric layers or nodes and edges emanate from a defined centre. For example, network diagrams use this structure.
Spatial		When displaying spatial analysis, the specific geographical areas and mapping projections used will determine the size and shape of the map structure. Values are plotted according to a longitude–latitude coordinate system or are associated with polygonal shapes of relevant geographic units.
Tabular		These structures are associated with table-like layouts based on associated x and y cell positions (like the heat map) or layouts that have different tiers or states (such as the Sankey diagram or the linear dendrogram).

Figure 10.14 List of Different Chart Shapes and Structures

'I'm obsessed with alignments. Sloppy label placement on final files causes my confidence in the designer to flag. What other details haven't been given full attention? Has the data been handled sloppily as well? … On the flip side, clean, layered, and logically built final files are a thing of beauty and my confidence in the designer, and their attention to detail, soars.' **Jen Christiansen, Graphics Editor at *Scientific American***

Elegant design: Like colour, composition decisions are always relative: an object's place and the space it occupies within a display create a relationship with everything else in the display. Unity in composition provides a similar sense of harmony and balance between all objects. The flow of content should feel logical and meaningful.

The enduring idea that elegance in design is most appreciated when it is absent is just as relevant with composition. Design solutions that felt effortless to navigate through visually will lead to a superior experience compared with those that felt punctured, chaotic and confusing. Thoroughness in the precision and consistency of your layout is important because any shortcomings will be immediately noticeable and will undermine the elegance. Pay attention to the smallest things: care about every last dot or pixel.

Summary: Composition

Features of Composition

This chapter explored the final element of developing your design solution concerning how you will organise the placement and sizing of all your visual elements within the space you have to work.

- Layout: What is the visual hierarchy of your project? Making decisions about the relative size and placement of all your visual elements, including charts, interactive controls and annotations.
- Arranging: Concerning the ordering and direction of your data content as it is displayed within a chart with different options including alphabetical, chronological, locational, ordinal and ranked sorting.
- Sizing: Ways of astutely using the technique of small multiples and correct approach to sizing charts through axis-scale ranges for different chart types.

Influencing factors and considerations

If these were the options, how did you make your choices? The influencing factors included:

- Medium: What space have you got to work within?
- Quantitative value range: What is the minimum and maximum value range and are there legitimate outliers (large or small) that will skew the distribution of values and create challenges for accommodating them?

- Editorial thinking: How many different angles (charts) might you need to include? Is there any specific hierarchy of importance or sequence that needs to be conveyed?
- Data representation: All charts have a spatial consequence and have varied structures and sizing requirements that will need to be accommodated.
- Elegant design: The unity of your layout, offering a seamless visual journey to the viewer, is another contributing factor that will create elegance in your work.

General Tips and Tactics

- Empty space is like punctuation in visual language – use it to break up content when it needs that momentary pause, just as a *comma* or *full stop* is needed in a sentence. Do not be afraid to use empty space more extensively across larger regions as a device to create impact. Like the notes not played in jazz, effective composition can be achieved through the distinction between something and nothing.
- At the deepest level of composition thinking you will become evermore consumed by the tiniest of precision judgements. The task of nudging things by fractions of a pixel and constantly resizing and realigning features will dominate your progress.
- As your energy starts to diminish, and your deadlines start to emerge, you will need to maintain a commitment to thoroughness and a pride in precision right through to the end!

What now? Visit book.visualisingdata.com

EXPLORE THE FIELD Expand your knowledge and reinforce your learning about working with data through this chapter's library of further reading, references, and tutorials.

TRY THIS YOURSELF Revise, reflect, and refine your skill and understanding about the challenges of working with data through these practical exercises.

SEE DATA VISUALISATION IN ACTION Get to grips with the nuances and intricacies of working with data in the real world by working through this next instalment in the narrative case study and see an additional extended example of data visualisation in practice. Follow along with Andy's video diary of the process and get direct insight into his thought processes, challenges, mistakes, and decisions along the way.

Epilogue

The Development Cycle

The previous five chapters have been leading you through design decision making, broadening your awareness of what options exist and then informing you about the things that will shape your choices. The consequence of this will be a fully reasoned design specification. This represents your conceptual thinking – a detailed plan of what you *intend* to develop.

In this closing section of this book, I want to leave you with an understanding of what happens next as you continue the process of developing your design (Figure E.1), translating your concept idea into a technically executed solution.

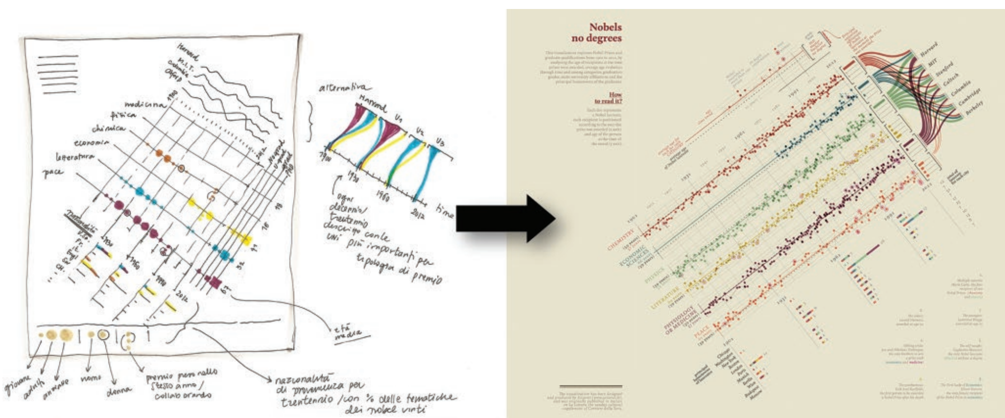
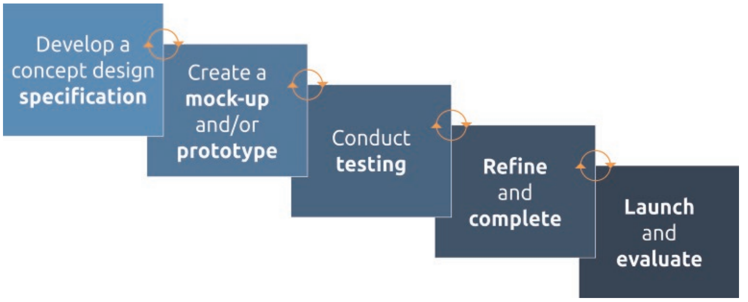


Figure E.1 From Concept to Solution: A Wireframe Sketch and Final Design for Nobels, No degrees, by Accurat

The development cycle (Figure E.2) is characterised by loops of iteration at the intersection of each successive step as you gradually nudge your ideas forward through successive rounds of enhancements. The degree of iteration required will depend on the size of the gap that exists between the idealism of your plan and the reality of what is technically achievable. Often, you will not really know what is feasible until you try it. Thoughtful planning can help restrict the size of this gap, but even then, the process of construction will likely incur impromptu revision or unforeseen compromises. Trade-offs between ambition and pragmatism are inevitable. At times, when things are just not progressing as you wish, you might need to go back to the

drawing board, returning to the start of your design thinking to conceive new choices and carve out a different path.

Figure E.2 The Stages of the Development Cycle



The relevance of and rigour required across each of these steps is for you to determine. If the solution you are creating is relatively simple in nature it will not likely involve the same demands of development as would be necessary with creating a complex interactive visualisation. You should see this cycle as an indicative outline; adapt it and use it as you see fit.

Let's learn more about the specific components of the development cycle. As mentioned, the consequence of the previous five chapters of this book led to completion of the first step, 'developing a concept design specification', which would have entailed creating initial storyboards and/or wireframes to capture what you intend to create. So what follows this?

Create a mock-up and/or prototype: Whereas wireframing and storyboarding are characterised by the creation of low-fi sketches, the development of *mock-ups* or *prototypes* advances the detail of your draft ideas towards a technical solution (assuming you are not producing your output by hand). This effectively leads to the creation of a first version that should closely demonstrate some of the main features of the eventual solution. With the next step in the cycle concerned with *testing*, this will understandably need to be focused on obtaining the most helpful feedback. The respective terms tend to be used interchangeably but I feel *mock-up* is more applicable for developing static work, whereas *prototype* is more relevant to interactive work.

Conduct testing: To move forward from a mock-up or prototype version requires testing. The first round of testing is done automatically and constantly by you and/or your collaborators to iron out any obvious immediate problems. In software development parlance, this would generally be consistent with *alpha testing*. Naturally, beta follows alpha and next you will need to seek others to 'use' it, evaluate it and give feedback on it. Even in small projects it is worth considering this, even if it involves offering the eventual viewer a first look. Testing happens regardless of the output format; it does not need to be a digital, interactive project to merit being tested. There will be different aspects of testing to conduct, and it is worth revisiting the three design principles to organise these:

- *Trustworthy* design testing concerns assessing the reliability of the work, in terms of the integrity of its content and performance. Are there any inaccuracies, mistakes or even deceptions? Are there any design choices that could lead to misunderstandings? Any aspects in how the data has been calculated or counted that could undermine trust? If it is a digital solution, what is the speed of loading and are there any technical bugs or errors? Is it suitably responsive and adaptable in its use across different platforms? Try out various user scenarios: multiple and concurrent users, real-time data, all data vs sample data, etc. Ask the people testing your solution to try to break it so you can find and resolve any problems now.
- *Accessible* design testing relates to how intuitive or sufficiently well explained the work is. Do viewers understand how to read it and what all the encodings mean? Is the viewer provided with a sufficient level of assistance that would be required in accordance with the defined characteristics of the intended audience? Can testers find answers to the questions *you* intended them to find and do this suitably quickly? Can they find answers to the questions *they* think are most relevant?
- *Elegant* design testing relates to questions such as: Is the solution suitably appealing in its design? Are there any features which are redundant or superfluous design choices that are impeding your engagement? Do you feel the appearance sustains any positive initial sentiment?

Whoever you invite to test your work will vary considerably in each project context. Generally, you may consider different cohorts of testers:

- *Stakeholders:* The ultimate customers/clients/colleagues who have commissioned the work may need to be included in this stage, if not for full testing then at least to engage them in receiving initial concept feedback.
- *Audience:* You might choose a small sample of your target audience and invite those viewers to take part in the beta testing.
- *Critical friends:* Peers/team/colleagues with suitable knowledge and appreciation about your working process may offer a more sophisticated means of expressing feedback.
- *You:* Sometimes (often) it may ultimately be down to you alone to undertake all testing, through either lack of access to other people or due simply to the lack of time. To accomplish this effectively you have to find a way to detach yourself from the mindset of the visualiser and try to occupy that of the viewer. You need to see the wood *and* the trees.

The timing of *when* to seek feedback through testing is another matter to consider. Sometimes the pressure from stakeholders requesting to see progress will determine this. Otherwise, you will need to judge carefully the right moment to do so. You do not want to get feedback when it is too late or change is expensive. Similarly, it can be risky showing

> 'We can kid ourselves that we are successful in what we "want" to achieve, but ultimately an external and critical audience is essential. Feedback comes in many forms; I seek it, listen to it, sniff it, touch it, taste it and respond.'
> **Kate McLean, Smellscape Mapper and Senior Lecturer Graphic Design**

nominated testers your undercooked concepts, perhaps just sharing early wireframes, when they might not have the capacity to imagine how this will materialise into a polished final solution.

Refine and complete: Based on the outcome of your testing process, this will likely trigger a need to address any issues that have been flagged or embrace new opportunities that emerge. Troubleshooting is one characteristic of this stage, as too is editing, which is more aligned with fine-tuning than problem solving. Regardless of the label, some of the features you are looking to cover here will include:

- correcting identified errors or issues;
- stripping away superfluous content;
- checking and enhancing the remaining content;
- squeezing out final degrees of sophistication from every layer of your design;
- improving the consistency and cohesion of your choices;
- double-checking the accuracy of every component;
- revisiting initial requirements and agreed definitions.

> 'You know you've achieved perfection in design, not when you have nothing more to add, but when you have nothing more to take away.' **Antoine de Saint-Exupéry, Writer, Poet, Aristocrat, Journalist and Pioneering Aviator**

In any creative process a visualiser is faced with having to declare work *complete*. Judging this can be quite a tough call to make. While the looming presence of a deadline (and, at times, agitated stakeholders) will sharpen the focus, often it comes down to a fingertip sense of when you feel you are entering the period of diminishing returns – when the refinements you make no longer add sufficient value for the amount of effort you invest in making them. Eventually, you will reach a judgement that your work is *good enough*. Completion is perhaps never truly reached in a creative process lacking *the* single perfect solution – you just need to finish.

> 'Admit that nothing you create on deadline will be perfect. However, it should never be wrong. I try to work by a motto my editor likes to say: No Heroics. Your code may not be beautiful, but if it works, it's good enough. A visualisation may not have every feature you could possibly want, but if it gets the message across and is useful to people, it's good enough. Being "good enough" is not an insult in journalism – it's a necessity.' **Lena Groeger, Science Journalist, Designer and Developer at ProPublica**

Launch and evaluate: The nature of launching your work will vary significantly, based, as always, on the context of your challenge. It might simply be emailing a chart to a colleague or presenting your work to an audience. In other cases, it could be a graphic going to print for a newspaper or involve an anxious go-live moment with the launch of a digital project. Whatever the context of your launch, there are a few characteristic matters to bear in mind. These will not be relevant to all project scenarios but, over time, you might encounter them in different situations:

- Are *you* ready? Regardless of the scope of your work, as soon as you declare work completed and published you are at the mercy of your decisions. You are no longer in control of how people will interpret your work and in what way they will truly use it. If you have a particularly large, diverse and potentially emotive subject matter, you will need to be ready for the questions and scrutiny that might head in your direction.
- *Communicating* your work is a big deal. The need to publicise and sell its benefits is of particular relevance if you have a public-facing project. You might promote it loudly or leave it as a 'slow-burner' to spread through word of mouth. For more modest or intimate audience types you might need to consider directly presenting your work to these groups, coaching them through what it offers. This is particularly necessary on those occasions when you may be using an unfamiliar representation approach.
- What ongoing *commitment* exists to support the work? This uniquely relates to digital projects. Do you have to maintain a live data feed? Will it need to sustain operations with variable concurrent visitors? What happens if it goes viral – have you got the necessary infrastructure? Have you got ongoing access to the people/skills required to keep the project alive and thriving?
- Will you need to revise, *update* and rerelease the project? Will you need to replicate this work on a repeated basis? What can you do to make the reproduction as seamless as possible?
- What is the work's likely *shelf life*? Does it have a point of expiry after which it could be archived or even killed off? How might you digitally preserve it beyond its useful lifespan?

There are two components in evaluating the outcome of a visualisation solution that will help to refine your capabilities: what was the *outcome* of the work; and how do you reflect on *your performance*?

Measuring the effectiveness of a data visualisation from an outcome perspective remains an elusive task. This is largely because it can only be determined according to contextual measures of success. This is why defining 'purpose' is necessary early on.

Sometimes effectiveness is tangible, but most times it is entirely intangible. If the purpose of the work is to further the debate about a subject, to establish reputation or voice of authority, these are hard things to measure in terms of positive outcome. One option may be to invert the measurement to seek evidence of tangible ineffectiveness. For example, there may be significant reputation-based impacts should decisions be made on inaccurate, misleading or inaccessible visual information.

There are, of course, relatively free quantitative measures available for digital projects, including web-based metrics such as visitor counts and social media engagement (e.g. likes, retweets, mentions). These, at least, provide a surface indicator of success in terms of the project's apparent appeal and spread. Ideally, however, you should aim to supplement this by collecting more reliable qualitative feedback, even if this can, at times, be rather expensive to secure. Some options include:

- capturing anecdotal evidence from comments submitted on a site, opinions attributed to tweets or other social media descriptors, feedback shared in emails or in person;

- informal feedback through polls or short surveys;
- formal case studies which might offer more structured interviews and observations about documented effects;
- experiments with controlled tasks/conditions and tracked performance measures.

A personal reflection of your contribution to a project is important for your own development. The best way to learn is by considering the things you enjoyed and/or did well (and do more of those things) and by identifying the things you did not enjoy/do well (and do less of those things or do them better!). Look back over your project experience and consider the following:

- Were you satisfied with your solution? If yes, why; if no, why and what would you do differently?
- In a different context, what other design solutions might you have considered?
- Were there any skill or knowledge shortcomings that restricted your process and/or solution?
- Are there aspects of this project that you might seek to recycle or reproduce in other projects? For instance, ideas that did not make the final cut but could be given new life in other challenges?
- How well did you use your time? Were there any activities on which you feel you spent too much time?

Developing effectiveness and efficiency in your data visualisation work will take time and will require your ongoing efforts to learn, apply, reflect and repeat again. I am still learning new things every day. It is a journey that never stops because data visualisation is a subject that has no ending.

'All of us who do creative work, we get into it because we have good taste … [but] there is this gap and for the first couple of years that you're making stuff, what you're making is just not that good … It's trying to be good, it has potential, but it's not. But your taste, the thing that got you into the game, is still killer. And your taste is why your work disappoints you. A lot of people never get past this phase, they quit. Most people I know who do interesting, creative work went through years of this. We know our work doesn't have this special thing that we want it to have. We all go through this. And if you are just starting out or you are still in this phase, you gotta know it's normal and the most important thing you can do is do a lot of work. Put yourself on a deadline so that every week you will finish one story. It is only by going through a volume of work that you will close that gap, and your work will be as good as your ambitions. And I took longer to figure out how to do this than anyone I've ever met. It's gonna take a while. It's normal to take a while. You've just gotta fight your way through.' **Ira Glass, Host and Producer of 'This American Life'.**

References

These references relate to content mentioned in the body text and/or attributed quotes that do not come from individual interviews with the author. Extensive further reading lists to support each chapter's content are provided in the companion digital resources.

Bertin, Jacques (2011) *Semiology of Graphics: Diagrams, Networks, Maps*. Redlands, CA: ESRI Press.

Chimero, Frank (2012) *The Shape of Design*. http://shapeofdesignbook.com/

Cleveland, William S. and McGill, Robert M. (1984) 'Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods'. *Journal of the American Statistical Association*, vol. *79*, no. 387, pp. 531–54.

Cox, Amanda (2013) *Harvard Business Review*. https://hbr.org/2013/03/power-of-visualizations-aha-moment/

Crawford, Kate (2013) 'The Hidden Biases in Big Data'. *Harvard Business Review*. http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html

de Bono, Edward (1985) *Six Thinking Hats*. New York: Little, Brown.

Glass, Ira (2009) Open Culture. www.openculture.com/2009/10/ira_glass_on_the_art_of_story_telling.html

Heer, Jeffrey and Shneiderman, Ben (2012) 'Interactive Dynamics for Visual Analysis'. *ACM Queue*, vol. *10*, no. 2, p. 30.

Ive, Jonny, Kemp, Klaus and Lovell, Sophie (2011) *Dieter Rams: As Little Design As Possible*. London: Phaidon Press.

Kahneman, Daniel (2011) *Thinking Fast and Slow*. New York: Farrar, Straus & Giroux.

Mackinlay, Jock (1986) 'Automating the Design of Graphical Presentations of Relational Information'. *ACM Transactions on Graphics (TOG)*, vol. *5*, no. 2, pp. 110–41.

Morton, Jill (2015) *Color Matters*. www.colormatters.com/color-and-design/basic-color-theory

Munzner , Tamara (2014) *Visualization Analysis and Design*. Boca Raton, FL: CRC Press.

Reichenstein, Oliver (2013) *Information Architects*. https://ia.net/know-how/learning-to-see

Rumsfeld, Donald (2002) *US DoD News Briefing*. https://en.wikipedia.org/wiki/There_are_known_knowns

Shneiderman, Ben (1996) 'The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations'. *Proceedings of the IEEE Symposium on Visual Languages*. Washington, DC: IEEE Computer Society Press, pp. 336–43.

Stefaner, Moritz (2014) *Well-formed Data*. http://well-formed-data.net/archives/1027/worlds-not-stories

Stevens, Stanley Smith (1946) 'On the Theory of Scales of Measurement'. *Science,* vol. *103*, no. 2684, pp. 677–680.

Tukey, John W. (1980) ‘We Need Both Exploratory and Confirmatory’. *American Statistician*, vol. *34*, no. 1, pp. 23–5.

Tversky, Barbara and Bauer Morrison, Julie (2002) ‘Animation: Can it facilitate?’. *International Journal of Human-Computer Studies*, Special issue: Interactive graphical communication, vol. *57*, no. 4, pp. 247–62.

Vitruvius Pollio, Marcus (15 BC) ‘De architectura’.

Wotton, Sir Henry (1624) *The Elements of Architecture*. London: Longmans, Green.

Index

In this index, titles or caption descriptions of visualisations are printed in italics. The numbers of the pages where they appear are in bold print. Numerals at the beginning of an entry are treated as words, e.g. ‘3D decoration’ is filed after ‘thoroughness’.

accessibility 45–50, 127, 243–4
  testing 293
Accurat 194, 236, 295
acquisition of data 71, 96–7, 115
adaptability 33
aesthetic seduction 53
aims of book 1–2, 5
Aisch, Gregor 126, 127
All the Buildings in Manhattan 197, **198**
alphabetical sorting 280–1
analysis and communication 6–7
analysts, qualities and traits of 112
Andrews, Wilson 79
angle of analysis 119–21, 123, 125, 126, 128
animating 217–19, 218, 227
annotating (interactivity) 215–17
annotations 19, 82, 125, 128, 231–47
  audience 244
  audio 242–43
  captions 240–1
  chart apparatus 238, 239
  chart references 238–9
  elegant design 246
  footnotes 243
  headings 231–3
  labels 240–1, 290
  and layout 280
  legends 237–8
  methods statements 243
  missing annotations 43
  and purpose 244–5
  reader guides 235–6
  and setting 244
  for specific chart types *see* charts, types, *gallery of charts*
  and text size 285
  user guides 233, 234–5
Annual Staff Perception Survey 261, **262**
API (Application Programming Interface) 96
appending data 109
areas of shapes 196
arranging 280–5
Art in the Age of Mechanical Reproduction: Walter Benjamin 267–70, **268**
Asia Loses its Sweet Tooth for Chocolate **53**
asking questions 116
‘at-a-glance’ viewing 76
attention span 49–50
attention to detail 36, 116, 247, 290, 291
attitudes 50
attributes (channels) 17, 18, 135, 137
audiences:
  and accessibility 45–50, 129
  attention span 49–50
  attitudes and emotions 50
  beta testing 295
  definition 10
  diversity of 31, 67
  empathy for 47
  and interactivity 227–8
  interests of 64
  knowledge of subject 67, 244
  language 67
  listening to 35
  motivation 67, 86
  needs 127
  and relevance 46, 64
  respect for 51–2
  thinking about in design process 33–4, 129
  and use of annotations 244
  and use of colour 273–4
  visualisation literacy 67, 244
  visualiser’s knowledge of 67
  see also complexity/simplicity; understanding
audio 240–41
‘Avengers’ characters’ appearance over time **179**
Average Weekly Brent Crude Oil Prices, 2008–2018 **175**
axes 12
axis scales 240–1, 285–7
axis titles 240

Baby Names in England and Wales 2017 **209**, 259
backups 116
Baldwin, Taylor 198
bandings 239
bar charts 21, 43, 75, 76, 140, 194
  axis scales 285, 286, 287
  truncation 286, 287

Barratt, John 238, 271
Baseball Home Run Trajectories **198**
Battling Infectious Diseases in the 20th Century: The Impact of Vaccines **240**
Baur, Dominik 81, 103, 282
Beccario, Cameron 221
Bees, Drew 126
Berkowitz, Bonnie 259
Berliner Morgenpost 215, 254
Bertin, Jacques: *Semiology Graphique* 190
bias 39
Black Students Are Underrepresented On Campus 82, **83**, **239**
Bloomburg Billionaires **152**
Bloomberg News 108
Bloomberg Visual Data 108
boardrooms 72
Boom and Bust: The Shape of a Roller-coaster Season **237**
Brady, Tom 126
brain 'states' 34–5
Breakdown of Michael Schumacher's F1 Career Over 308 Races **157**
Breathing Earth 218, **219**
Bremer, Nadieh 219
brief: formulating 32, 61
brushing 210
Bryant, Kris 197, 198
bubble plots 167, 207
budget 70
Bui, Quoctrung 124
bump charts 172, 209, 284
Burn-Murdoch, John 265
Buying Power: The Families Funding the 2016 Presidential Election 77, **79**, 89

Cable, Dustin A. 221
Cairo, Alberto 62
Camoes, Jorge 37
Campbell, Sarah 227
captions 240–1
Carli, Luis 85
Carbon Map **186**
cartesian charts 291
Casualties **272**
Cesal, Amy 73
Chan, C. 90
Chang, Alvin 224, 258
channels *see* attributes (channels)
chart apparatus 238
chart references 238–9
Charting the Beatles: Song Structure **260**
Chartmaker Directory **189**–90
charts:
  and *graphs, plots* and *diagrams* 12
  lines and attributes 17–18
  and tables 19
  technological constraints 189–90
  3D form 43
  types 11–12, 18
    'CHRTS' family of types 138
    gallery of charts 138–88
      area cartogram 186
      area chart 175
      bar chart 140
      beeswarm plot 153
      box-and-whisker plot 156
      bubble plot 167
      bullet chart 142
      bump chart 172
      chord diagram 170
      choropleth map 180
      clustered bar chart 141
      connected dot plot 146
      connected scatter plot 174
      dendogram 164
      density plot 155
      diverging bar chart 160
      Dorling cartogram 187
      dot map 184
      dot plot 152
      flow map 185
      Gantt chart 178
      grid map 188
      heat map 150
      histogram 154
      instance chart 179
      isarithmic map 181
      line chart 171
      Marimekko chart 161
      matrix chart 151
      network diagram 168
      pictogram 147
      pie chart 157
      polar chart 145
      prism map 183
      proportional symbol chart 148
      proportional symbol map 182
      radar chart 144
      Sankey diagram 169
      scatter plot 166
      slope graph 173
      stacked area chart 176
      stacked bar chart 159
      stream graph 177
      sunburst chart 163
      treemap 162
      Venn diagram 165
      Voronoi treemap 138
      waffle chart 158
      waterfall chart 143
      word cloud 149
    and range of values 194
    and types of data 194, 195
  unfamiliar types 49
  usage 43

Cheshire, James 238, 271
Chimero, Frank, *The Shape of Design* 51
choices:
of chart types 80
and restrictions 66
what to include/exclude 122
see also decision making
chord diagrams 170, 284
Chow, Emily 259
Christiansen, Jen 54, 290
Chrome Dominates a Cluttered Browser Market **195**
chronological sorting 281
City of Anarchy 277–9, **278**
Ciuccarelli, Paolo 190
clarification 49
cleaning data 106
Cleveland, William and McGill, Robert: 'Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods' 190, 192
Clever, Thomas 52, 108
CNN 271
cockpit setting 72
coffee shops 72
collaboration 69
Collins, Keith 108
Colors of the Rails 260, **261**
colour 247–74
applications 19, 125, 126, 127
background neutral tones 270
categorical classifications 256–65
categories of colour 260
with large numbers of categories 259–60
CIELAB model 252
CMYK (Cyan, Magenta, Yellow, Black) model 250
contrasting 82, 253
converging/diverging colour scales 254, 255
functional decoration 265–72
harmony 267
HSL (Hue, Saturation, Lightness) model 250–2
impact of 45
judging variation 192
legibility 252
maps 266
medium 272
print quality 275
and purpose 273
quantitative scales 253–6, 254, 255
rainbow scale 255–6
RGB (Red, Blue, Green) model 250
rules and guidelines 272–73
and visual impairments 273–4
colour-blind-friendly altenative tones 274
commitment 297
communication 35–6, 114
comparing circles 196
Comparing the Degree of Trust by Australians for Different Institutions **159**
Comparing Relative Values and Daily Changes of Market Capital for Stocks **160**
completing 296
complexity/simplicity 46, 48–9
in representation 49
complicated subjects 48
composition 19, 129, 277–92
arranging 280–85
and data representation 290
and data types 99
and editorial thinking 125, 289–90
and elegance 51, 290
features 291
layout 277–9, 288
and medium 288
quantitative value range 288, 289
sizing 285–7
for specific chart types *see* charts, types, *gallery of charts*
see also style
comprehending (phase of understanding) 24–5, 74
confusion 47, 48, 49
connected scatter plots 174, 217
consolidating data 108–9
constraints 66, 69–71
consuming: definition 11
content 86
context:
circumstances 65–6
constraints 69–71
and content 86
deliverables 71–3
motivating curiosity 62–5
people 66–9
Coral Cities 285, **286**
correlation 12
Corum, Jonathan 269
Countries With The Most Land Neighbours **140**
Cox, Amanda 79
Crawford, Kate 42
creative influences 70
creativity 33, 36, 53
Cricketer Alastair Cook Plays His 161st Final Test Match 264, **265**
critical thinking 5
Cruz, Peter 193
Cultural Politics: Marijuana Use and Same-sex Marriage in USA **165**
cultural sensitivities 70
curiosity 62–5, 74
multiple curiosities 64–5
and question forming 65, 119
sample statements 65
specificity 64
customers *see* stakeholders

Daily Indego Bike Share Usage in Philadelphia **176**
Daily Mail 44, 45
Daily Price and Availability of Super Bowl Tickets **174**
dashboards 28
data:
  creating 106–7
  perfect data 115
  quality/condition 105, 131
  security/privacy 116
  technologies for presentation 71
  types 11, 16, 97–102
  working with 32, 96–115
    acquisition 71, 96–7, 115
    examination 71, 97–105, 115
    exploration 109–115
    sorting 280–85
    transformation 71, 106–9, 115
data art 28
data journalism (data-driven journalism (DDJ) 28
data representation 133–99
  chart types 11–12, 18
    gallery of charts 136–86
  and composition 288
  influenced by angle and framing 124–5, 128
  interactivity 226
  technological constraints 71, 189–90
  trustworthy design 196–200
  visual encoding 135–7, 136
data science 28
data source: definition 11
data visualisation: definition 15
datasets 11
  cross-tabulated 98–9
  expanding/appending 108–9
  normalised 98
de Bono, Edward, *Six Thinking Hats* 67
De Niro, Robert 63
deadlines 69
Deal, Michael 260
decision making 20, 31, 39, 118
  optimising 37–8
decoration 52–5
  3D decoration 197
  see also functional decoration
deductive reasoning 113
D'Efilippo, Valentina 92, 241
deliverables 71–3
  interactivity 225
DeSantis, Alicia 79
design characteristics 74
design principles 37–56, 38, 57
  accessibility 45–50, 245–6
  elegance 50–56, 290
  trustworthiness 38–45, 196–200, 226, 227
design process 31–7, 57
  process and procedure 33
  reasons for following 32–7
  stages 32
  see also design principles
design restrictions 70
development cycle 291–6
  stages 292
diagrams: definition 12
'Dialect quiz map' 85, 86
differences in projects 32–3
digital resources 10, 225
Dimensional Changes in Wood **85**
discrete/continuous data 102
Do You Remember Where Germany Was Divided? **215**
documenting 35
doing (practical undertakings) 34
  and not doing 36
domain expertise 97
donut charts 194, 195
Doping under the Microscope **288**
Du Bois, W.E.B. 6
duration of task 69–70

Earth 217, **219**
Ebb and Flow of Movies: Box Office Receipts 1986–2008 **177**
ECB Bank Test Results 211, **211**
Ecological Footprint and Biocapacity **167**
editing 51, 119
editorial angle 196
editorial thinking 119–31
  and composition 125, 289–91
  establishing 32
  examples:
    Fall and Rise of US Inequality 123–5, **124**
    Why Peyton Manning's Record Will Be Hard to Beat 125–9, **126**, **127**
effectiveness 297, 298
elegance 50–56
  in annotations 246
  and composition 51, 290
  interactivity 226–7
  testing 293
eliminating the arbitrary 51
Elliot, Kennedy 33, 217, 234
emotions 26, 50, 76–80, 190
  manipulation of 79
emphases 122
empty space 292
enclosure charts 291
enlightenment 26
environmentally friendly design 56
Equal Earth projection 199
ER Wait Watcher: Which Emergency Room Will See You the Fastest **281**

erroneous values 105
evaluation 297–8
Evans, Tom 213
Every Time Ford and Kavanaugh Dodged a Question 257–**8**
examination of data 71, 97–105, 115
Excel 99
Executive Pay by the Numbers **256**–7
exhibitory visualisations 86–8, 245
expanding data 108–9
experiences offered by visualisation 82–8
experimentation 33
explanatory visualisations 82–4, 87, 245
explorable explanations 86
exploration of data 109–115, 195–6
exploratory data analysis (EDA) 111–15
exploratory visualisations 84–6, 245

facilitating 26, 37
Fall and Rise of US Inequality, The 123–5, **124**
Falling Number of Young Homeowners **44**
familiarity 49
feedback 295–6
  from friends 295
feeling tone 76–81, 190
figure-ground perception 40
Filmographics 62–**3**, **279**
filtering 202–5
Financial Times 67, 97, 265
FinViz: Standard & Poor's 500 Index Stocks **80**, **207**, **274**
FiveThirtyEight 83, 239, 284, 290
Flasseur, Vincent 212
focus 122–23, 124, 126, 127, 128
fonts 245, 246
Foo, F. 90
footnotes 243
foraging for data 97
Forbes: The World's 100 Highest-paid Athletes 87, **88**, **147**
Ford, Christine Blasey 257
Forecast % Chance of Winning Presidency (US Election, 8th November 2016) **25**–6
formats 12
formatting data 108
Four Teams in Group F of 2018 World Cup **145**
framing 122, 123, 126, 128
frequency 73
frequency counts 103
frequency distribution 103
Frequency of Words Used in Ch 1 of First Edition of This Book **149**
Fuller, Richard Buckminster 51
functional decoration 266–72
functionality 12
Funds Raised Across USA for Election Candidate Hillary Clinton **182**
Funke Interaktiv 215

Gender Pay Gap US/UK **146**
geometric miscalculations 196
German General Election Results Showing Winning Party for Each Location **180**
Ghael, Avni 193
Glass, Ira 298
Global Flow of People **170**
goal of visualisation 74
  see also purpose
Goddemeyer, Daniel 282
Goldsberry, Kirk 79
Gourlay, Colin 214
Grabell, Michael 210
Grape Expectations **90**
Graphic Language: The Curse of the CEO 107–**8**
graphics 12
graphs 12
  see also charts
Gray, Marcia 106
Green, Jeff 108
Grimwade, John 222
Groeger, Lena 46, 281, 296
Grothjan, Evan 79
Growth in Participants and Female Participation at the Summer Olympics **161**
Gun Deaths in Florida **40**, 42, 79
Guys Named John, and Gender Inequality **263**

harmony 265
harnessing ideas 88–93
headings/titles:
  artistic 232–33
  descriptive 232
  and introductions 233, 234
  as questions 232
  statements 231–33
  and storyboarding 280
Heer, Jeff and Shneiderman, Ben: 'Interactive Dynamics for Visual Analysis' 226
Hemingway, Ernest 29
hidden thinking 32
Highest Max Temperatures in Australia **255**–6
History Through the President's Words 216, **217**, **233**
Hobbs, Amanda 36, 48
Holdouts Find Cheapest Super Bowl Tickets Late in the Game 216, **218**
Holmes, Nigel 91
honesty 38, 39
Horse in Motion **229**

Household Incomes for Simulated Population of Chicago Residents **153**
Housing and Home Ownership in the UK **44**, 45
How the 'Avengers' Line-up Has Changed over the Years **216**
How Big Will the UK Population Be in 25 Years' Time? **210**
How Each State Generates Electric Power (2004–2014) **173**
How Good is 'Good'? **155**
How Inclusive are Beauty Brands Around the World? **154**
How Long Will We Live – And How Well? **166**, **259**
How Nations Fare in PhDs by Sex 204, **206**
How Popular is Your Birthday? **150**
How Well Am I Running? **121**, 122
How Well Do You Know Your Area? 212, **213**
How Y'all, Youse and You Guys Talk **86**, **181**
Hubley, Jill 206
hue 250–52
Hurt, Alyson 69, 113, 114

If Vienna Would Be an Apartment **261**, 261
inductive reasoning 113
info-posters 27
infographics 27
information design 27–8
information visualisation 27
Ingold, David 108
innovative design 55
integrity 43–5, 295
interactivity 125, 128, 203–230
  accessibility 227–8
  animating 217–19, 218
  annotating 215–17
  with colour for categorical classification 259
  data representation 226
  elegance 228–9
  event, control, function 204
  in explorative visualisations 84, 85
  filtering 204–7
  highlighting 207–11, 208
  influencing factors 225–9, 226
  and layout 280
  and medium 288
  navigating 219–24, 220
  participating 211–215, 212
  technologies for 71, 225
  user guides 233, 234–5
interestingness 129
internet as source of data 96
interpreting 22–3
interval data 101
introductions 233, 234
Iraq's Bloody Toll **41**, 42
iteration 33, 130
Ito World 286

Jacobs, Brian 87, 289
Jenkins, Nicholas 235
Johnson, Richard 217, 234
judging comparative size 191

Kahneman, Daniel, *Thinking Fast and Slow* 88–9
Kasich Could Be the GOP's Moderate Backstop **284**–5
Katz, Josh 86
Kavanaugh, Brett 257
Keegan, Jon 214
Killing the Colorado: Explore the Robot River **222**
Kindred Britain **235**
Kirchner, Lauren 222
Kirk, Andy 63, 237, 243, 257, 279
Klack, Moritz 254
Klee, Paul 34
Knott, Matt 63, 279
Known Knowns **111**, 113
Kocinova, Lucia 92, 241

labels 240–1, 290
Lambert Arimuthel Equal-area projection 199
Lambrechts, Maarten 283
Larson, Jeff 222
launching 296–7
layout 277–80
  and annotations 280
  and interactivity 280
  and size 279
learning 114, 116
legends 12
legibility 246
  colour 252
levels of data 99
Li, Jason 154
Liberals Most Likely to Favor No Restrictions on Abortion **160**
Life Cycle of Ideas **236**
lightness (colour) 251, 253
Lindemann, Todd 259
line charts 18, 239, 284
  aspect ratio 43
  axis values 276, 287
linking 208
Lionel Messi: Games and Goals for FC Barcelona 20–22, **21**
locational sorting 281–2
log scales 289, 290
long-lasting data 56
long-lasting design 55–6
Losing Ground 86, **87**, 288, **289**
Lunge Feeding **269**–71
Lupi, Giorgia 91, 194
Lustgarten, Abraham 222

McCandles, David 213
McGill, Robert 190
Mackinlay, Jock: 'Automating the Design of Graphical Presentations of Relational Information' 190, 191
McLean, Kate 100, 295
making 34
Making Sense of Skills: A UK Skills Taxonomy **234**
Manian, Divya 154
Manley, Ed 238, 271
Manning, Peyton 125, 126, 127, 128
Manovich, Lev 282
maps 12, 291
  and colour 266
  projection 197, 198–200
  thematic 43, 197, 198–200, 199
  see also charts
markers 239
Market Capitalisation of Companies **148**
market influences 70
market share browsers 194, 195
marks 17–18, 135, 136
Marshall, Bob 87, 289
measurement of central tendency 103
measurements of spread 103
media (formats) 72–3
  and colour 272
  and composition 288
  and interactivity 288
Mediafin 283
Meirelles, Isabel 47
Mellnik, Ted 217, 234
memorability 56
Mercator projection 199
Mercer 'Quality of Living City' 285
MeTooMentum **92**, **241**
Migliozzi, Blacki 223
minimalism 51
minimum friction 47
missing values 105
mistakes 36, 43
  and geometric miscalculations 196
Mizzou's Racial Gap is Typical on College Campuses 82, **83**
mock-ups 294
Mollweide projection 199
Month in an Animal Shelter **169,** 226, **227**
Morrison, Julie Baur 228
Morton, Jill 267
multiple media production 72
Munzer, Tamara 285
Murray, Scott 66, 235
Muybridge, Eadweard 228

narrative visualisations 83–4
National Geographic 33
Native and New Berliners – How the S-Bahn Ring Divides the City **254**
navigating 219–24, 220
Nearly Half of New Zealand's Migration Gain is From Asia **143**
Nelson, John 79, 116
New York City Street Trees by Species 204, 205, **206**, 259
New York Times 79, 85, 86, 112, 126, 127, 253, 256, 263, 269
news and media organisation: style 52
NFL Players: Height and Weight over Time 219, **220**
Nightingale, Florence 6
Nobel Laureates 207, **208**
Nobel Laureates Awarded (1901-2017) by Country of Birth **75**–6
Nobel Laureates by Category and Country of Birth **151**
Nobels, No Degrees **295**
NOIR (TNOIR) classification 99–100, 115
nominal data 100
Noonan, Laura 212
note-taking 35, 94, 116
nothings (in data) 114
Number of Vehicles Using Hong Kong's Network of Roads 2011 **185**
NYPD Staffing Compared With Other Cities **264**
NZZ 252, 261

Obama's Health Law: Who Was Helped Most? **253**, 254
objectivity 39
O'Brien, Oliver 238, 271
observation-meaning gap 23
OECD Better Life Index **81**, **103**, 192
  Ireland **224**
Office for National Statistics (ONS) 44, 45, 97, 209, 210, 213
On Broadway **282**
One Angry Bird 76, **77**
100 Years of Tax Brackets, in One Chart 223, **224**
opinion 120
ordinal data 101
ordinal sorting 282, 283
orientation of charts 284–5
Ortiz, Santiago 88
outliers 12

Parshina-Kottas, Yuliya 79
participating 211–215, 212
Peek, Katie 35, 244
perceiving (phase of understanding) 21–2
Percentage of Hours During 2017 the Sun was Above the Horizon in Nuorgam, Finland **158**
perceptual accuracy 191–93

perfection, impossibility of 31
Periscopic 77, 206, 211, 242
pertinence 129
photography 120
physical displays 197, 198
pie charts 157, 194, 195
plagiarism 89, 91
Playfair, William 6
plots 12
  see also charts
political pressures 70
Politically Important Topics for Germans Between 1998 and 2017 **172**
Politizane 84, 242
Pong, Jane 72, 91, 277
Popularity of International Outlets 72, **73**
Posavec, Stephanie 33, 51, 268
presentation 19, 50
pressures 70
primary data 97
print as medium 72, 288
project circumstances 65–6
project management 230
projects: definition 11
Proportion of Sales Percentage by Channel over Time **16**–17, **18**, 24–5
ProPublica 87, 210, 222, 281, 288, 289
prototypes 294
publicising work 297
publishing 71
purpose 61, 74, 76
  and annotations 244–5
  and choice of charts 190–4
  and colour 273
  and interactivity 225, 226
Pursuit of the Faster **257**
Pursuit of the Faster (footnotes) **205**, 243
Pyensen, Nicholas D. 269

Qiu, Yui 210
quantity 73
Quealy, Kevin 126, 127, 253
question forming 65

Racial Dot Map, The **222**
radial structures 291
radio buttons 207
Rain Patterns **281**
Rams, Dieter: Principles of Good Design 37, 38, 45, 50, 55
ranked sorting 282
Ranking the Ivies **156**
Ranking of Perceptual Tasks **191**
Rapp, Bill 71
ratio data 101
Raureif GmbH 81, 103
raw data 11
Razor Sales Move Online, Away from Gillette 53, **54**
reading tone 75–6, 190, 245
reasoning 112–13
recurrence of concern 33–4
reducing randomness of approach 32
reference lines 239
refining 296
reflective learning 36–7
Reichenstein, Oliver 34, 52
relevance 46–7, 129–30, 230
representation complexity 49
research 36, 114
Reuters Graphics 90, 210, 287
Rim Fire – The Extent of Fire in the Sierra Nevada Range and Yosemite National Park 260, **261**
Ring-Necked Parakeets 266–**7**
Roberts, Graham 79
Roston, Eric 223
Rougeux, Nicholas 236, 261
Rumsfeld, Donald 110
Runs Scored in Test Matches by English Batsmen **171**
Russell, Karl 256
Ryan, Claudine 214

Saint-Expuéry, Antoine de 296
samples 105
Sandler, Adam 63
Sanger-Katz, Margot 253
Sankey diagrams 167, 284
sans-serif typefaces 245
saturation 251
scales 12
scales of measurement 99
Scarr, Simon 41, 90, 121, 278
scatter plots 82, 83, 125, 239
Scientific American 54, 290
scientific visualisation 28
scrapbooks 89, 94
'scrollytelling' 222
series 11
serif typefaces 245
settings 71–2
'Seven Hats of Visualisation Design' 68
shape of data 103–4
Share of Individuals Using the Internet 2015 **187**
Share of People Voting to Leave and Remain During the EU referendum **188**
Shaw, Al 87, 222, 289
Shibuya, Felipe 193
shifting focus 64
Shneiderman, Ben 80, 226
Simmon, Robert 263
simplifying 48–9
size 277, 283–5
  and quantitative value range 288, 289
  restrictions 70
  shrinking 285

size of data 102
sketching 35, 91, 92
skills needed for this book 3
skills of visualisers 8, 67–9
Sleeman, Cath 234
Slobin, Sarah 56, 127
small mulitples 285, 286
Smith, Alan 67, 97
Snow, John 6
Songs That Were #1 in the UK Charts for the Greatest Number of Weeks 282, **283**
sophistication of content 46
sorting data 209, 280–5
sources of data *see* data, acquisition
South China Morning Post 41, 278, 281
Sparkes, Sophie 267
spatial analysis 291
Spielberg, Steven 63
  Jaws 55
Spotlight on Profitability **104**
Spraggon, Ben 214
squint test 273
stakeholders 63–4, 66, 295
Stamen 271
'Stand Your Ground' law (2005) 40
statistics:
  knowledge of 7
  supplemented with visuals 111–12
Stefaner, Moritz 81, 103, 104, 111, 120, 282
Stevens, Stanley 99
storyboarding 279, 280, 294
storytelling 28–9, 222–23
Streep, Meryl 63
style 52, 53, 70
  for colour usage 272–3, 275
subject distinctions 27–9
subject-matter experts 66
subjectivity 39
sufficiency 129
Sugar Quiz: How Much Sugar Is in Our Food? **214**
SUL-CDR 235
supplied data 96
Swift, Taylor 76, 78
System 1 and System 2 thinking 88–9
system download 96
Szücs, Krisztina 104

tabular designs 291
tabulation of data 11, 16–17, 24–5, 98–9
  NOIR (TNOIR) classification 99–100
Tabuleau 99
Tallest Buildings Around the World (Effect of Truncation on Bar Charts) 286, **287**
Taylor, Craig 286
Taylor Swift is Mostly Happy, Quite Often Sad, Sometimes Mad, and Occasionally Really Scared 76, 77, **78**
technological constraints 71, 189–90
  interactivity 223
Ten Actors Who Have Received the Most Oscar Nominations for Acting **141**
testing annotations 247
testing ideas 294–6
text size 285
textual data 100, 107–8
The Lens 87, 289
thinking 34
Thomas, Amber 154
thoroughness 51–2
3D decoration 197
Tigas, Mike 281
time 49–50, 118
time management 34
time-based data 102
timeliness 129
timescales 69–70
tone 74, 75–81, 190
  choice of 75, 80
tones (colour) *see* saturation
Top 20 Ranked Batters in Men's Test Cricket **142**
Total Sightings of Winglets and Spungles **23**
transformation of data 71, 106–9, 115
Transport for London (TfL) 96
Tree for US Immigration, A **193**, 194
treemaps 80
Trillions of Trees **183**
Tröger, Julius 254
troubleshooting 296
trust and truth 39–40
trustworthiness 38–45
  data acquisition 96
  in design 196–200, 226, 227
  integrity 43–5
  testing 295
Tufte, Edward 54
Tukey, John 112
Tulp, Jan Willem 109
Tversky, Barbara and Morrison, Julie Baur: *Animation: Can it Facilitate* 228
Twitter NYC: A Multilingual Social City **238**, 270–**1**
200+ Beer Brands of SAB and AB InBev 282, **283**
typefaces 245, 246
typology 245–6

Ulmanu, Monica 210
understanding 47–50, 67, 74
  and confusion 47, 48, 49
  content and context 86
  delivering 26
  facilitating 37, 74
  phases 20–25, 74
United Kingdom: Global Competitiveness **144**
UK Skills Taxonomy **164**

univariate/bivariate/multivariate techniques 7
University of Missouri 82
unknowns 111, 113
updating projects 297
Upshot, The 79
US Guns Deaths **211**, 241, **242**
US National Parks **178**, **236**
US residents Based on Location in 2010 Census **184**
usefulness 129

value intervals 285
value labels 241
variables 98
Veltman, Noah 220
Viégas, Fernanda 266
viewers:
  definition 10
  diversity of 26
  knowledge 22, 23, 24
vision: definition 74
visual analytics 28
Visual Cinnamon 219
visual encoding 135–7
  attributes (channels) 17, 18, 135, 137
  marks 17–18, 135, 136
visual immediacy 192, 194
visualisation literacy 67, 190
visualisation as prop 72
visualisers 67
  definition 10
  as leader 92–3
  'Seven Hats' (skills) 67–9, 68
visualising data 111
Vitruvius Pollio, Marcus, 'De architectura' 38, 45, 50
Voting Patterns for Democrats and Republicans Across Members of US House of Representatives **168**

Walker, Jonni 265
Wall Street Journal 53, 54, 216, 240, 264
Washington Post 217, 234
waterfall charts 143, 286
Watkins, Derek 79
Wattenberg, Martin 266
Wealth Inequality in America **84**, **242**, 243
web crawling 96
web scraping 96
Weber, Matthew 208
Wei, Sisi 281
Wendler, David 254
What are the Current Electricity Prices in Switzerland? **252**, 253
What Good Marathons and Bad Investments Have in Common 111, **112**
What's Really Warming the World? **223**
'Where's Wally?' 113
Which Companies Caused Global Warming? **163**
Who Old Are You? 212, **213**, 214
Why Peyton Manning's Record Will Be Hard to Beat 125–9, **126**, **127**
Wihbey, John 193
Wind Map 265, **266**
wine industry 89, 90
Winkel-Tripel projection 199
wireframing 279–80, 294
Witherley, Andrew 243, 257
Wolfers, Justin 112, 263
Workers' Compensation Reforms by State 209, **210**
World Top Incomes Database 124
Worst Games Ever Made, The 289, **290**
Wotton, Sir Henry 38
Wu, Shirley 78

x-axis 12, 284, 289

y-axis 12, 42, 284
Yourish, Karen 79

Zamora, Amanda 222
zooming 221